Leo Gürtler
Günter L. Huber

Combinación de Métodos
CUAL, CUAN Y LÓGICA
Puntos de vista objetivos y
Observaciones subjetivas

Este libro se ha realizado sin la amable ayuda de nadie.

Autores y contacto:

Dr. Leo Gürtler, osf.io/69gfz, github.com/abcnorio
Prof. Dr. Dr. h.c. Günter L. Huber, www.aquad.de

1ª Edición 2023 (2023-12-11)

Composición tipográfica: Leo Gürtler
Esta publicación se ha maquetado con LYX y LATEX2" (clase memoir) en Linux Debian Buster o Bullseye.
Los análisis de datos se realizaron con la ayuda de R, JAGS, BUGS y AQUAD 7 y 8.

Licencia CC BY-NS-SA 4.0

Esta edición del libro está disponible digitalmente como libro electrónico (pdf) en varios sitios web. El Rcode y los conjuntos de datos accesibles bajo licencia están disponibles en https://github.com/abcnorio/mixedmethod-rcode o https://osdn.net/projects/mixedmethod-rcode/. Allí encontrará hash MD5 del presente libro v.2023-10-23. Los futuros cambios en el libro, como correcciones de errores actualizaciones, etc. darán lugar a cambios en la versión y el hash MD5. El código R en sí está sujeto a la licencia GPL v3 (https://www.r-project.org/Licenses/AGPL-3) con la excepción de código adoptado y modificado de otros autores – en estos casos se aplican sus licencias.

Para todos aquellos
que hacen su trabajo
a disposición de los demás
sin esperar nada a cambio

Índice

Prólogos 15
Agradecimientos 19
Estructura del libro 21

Parte I – Comprensión y Sabiduría

Capítulo 1: *Discurso Precursor de la Teoría de Ciencia* **29**
1.1 El mundo es relativo 29
1.2 Conocimiento, comprensión y sabiduría 30
1.3 Los conocimientos dependen de la cultura 31

Capítulo 2: *Razonamiento Lógico en el Proceso de Investigación* **35**
2.1 La conclusión inductiva 37
2.2 La conclusión deductiva 48
2.3 La conclusión abductiva 52
2.4 ¿Se aplica la trinidad al razonamiento científico? 54
2.5 Excursus – Jugar, hacer y fabricar 57

Capítulo 3: *Hitos de la Filosofía de la Ciencia* **63**
3.1 Hume 64
3.2 Popper 64
3.3 Lakatos 65
3.4 Kuhn 65
3.5 Feyerabend 66
3.6 Conclusión – Filosofía de la ciencia 67

Parte II – Métodos Cuantitativos

Capítulo 4: *La Estadística Clásica* **71**
4.1 Un comienzo 71
4.2 Esquema de las teorías estadísticas 73
4.2.1El concepto de probabilidad. 73
4.2.1.1 Expectativas y subjetividad de las probabilidades 75
4.2.2 Significado y finalidad de las estadísticas 77
4.2.3 Estadísticas no concluyentes 79
4.2.3.1 Estadísticas descriptivas 79
4.2.3.2 Estadísticas exploratorias 79
4.2.4 Formas de estadística inferencial. 79
4.2.4.1 Estadísticas frecuenciales 80
4.2.4.2 Estadística de Bayes 80
4.2.5 La prueba estadística 81
4.2.6 Métodos mixtos 83
4.2.7 Ajustar las estadísticas 84
4.2.8 Otras lecturas y programas informáticos 85
4.3 Estadísticas clásicas. 86

4.3.1 La competencia en la estadística clásica: Fisher frente a Neyman-Pearson 86
4.3.2 La teoría de R. A. Fisher 86
4.3.2.1 Beber té y reconocer la leche 88
4.3.3 La teoría de Neyman-Pearson 91
4.3.3.1 Relevancia práctica 92
4.3.3.2 Dirección y tamaño: dos tipos de error subestimados 98
4.3.3.3 Modelos y respuestas - otros dos tipos de error 104
4.3.3.4 Estudio de caso Neyman-Pearson: gestión de la calidad. 105
4.3.3.5 Contabilidad de costes totales. 110
4.3.3.6 Intervalos de confianza 113
4.3.4 Pruebas y estimaciones - Confianza y valor p 128
4.3.5 Excursus - Simulaciones 130
4.3.5.1 Simulación de caso - Coleccionables de fútbol 133
4.3.6 Tamaño de la muestra 140
4.3.7 La subjetividad en la estadística clásica 145
4.3.8 El procedimiento de una prueba estadística: el ritual nulo 147
4.3.9 La probabilidad de los datos como base de las decisiones sobre las pruebas 151
4.3.9.1 Cálculo del valor p. 153
4.3.9.2 Sobre la relación entre la distribución normal y la distribución t 160
4.3.9.3 Excursus - Teorema del límite central 162
4.3.10 Epistemología recargada: la estadística clásica 168
4.4 Fuentes de error y fenómenos estadísticos 178
4.4.1 Distorsiones generales 179
4.4.2 En busca de la significación: intenciones de investigación inconscientes 181
4.4.2.1 El estudio de "Fifty Shades of Gray" 183
4.4.2.2 El estudio de Bem sobre la clarividencia 183
4.4.3 La gestión del poder. 189
4.4.3.1 Ejemplo de investigación: Simulación de potencia - WELL WELL 191
4.4.4 Réplicaciones 203
4.4.4.1 Índice R 204
4.4.4.2 Curvas Z 207
4.4.5 (Auto)Engaños 212
4.4.6 Estimación insesgada de las varianzas muestrales 212
4.4.7 Aleatorización. 217
4.4.8 Datos faltantes 226
4.4.9 Procedimientos equivalentes 239
4.4.9.1 Equivalencia de los procedimientos y métodos de medición 239
4.4.9.2 Equivalencia de los procedimientos y métodos de análisis de datos 254
4.4.10 Distribución normal de los residuos 256
4.4.11 Homogeneidad de la varianza (homocedasticidad) 261
4.4.12 Valores atípicos y datos de entrada elevados 262
4.4.13 Tamaño del efecto, frecuencia y relación con la escala original 270
4.4.14 Paradojas en la estadística 278
4.4.14.1 La paradoja de Simpson 278
4.4.14.2 La paradoja de Jeffreys-Lindley 282
4.4.14.3 La paradoja de Meehl 287
4.5 Conclusión - Resumen de Fisher vs. Neyman-Pearson 288

Capítulo 5: *Análisis Exploratorio de Datos (AED) según Tukey* **293**
5.1 Encontrar, buscar y revisar estructuras 293
5.2 Procedimientos típicos de AED en R 293
5.3 El análisis robusto de datos como parte de AED. 296
5.4 Diferenciación de AED de los enfoques confirmatorios 305
5.5 Estudios de caso de AED 306

5.5.1 Comparación de la población en los Estados federados de Alemania 307
5.5.2 Fertilidad y fecundidad 313
5.5.3 Distinción de especies en biología. 322
5.5.4 Vivir y morir en el Titanic 327
5.5.4.1 Contexto 328
5.5.4.2 Cuestionamiento 328
5.5.4.3 Reconstrucción del contexto histórico. 329
5.5.4.4 Datos sobre el Titanic 329
5.5.4.4.1 Capacidad y equipamiento 329
5.5.4.4.2 Proceso de hundimiento 330
5.5.4.3 Tiempo. 330
5.5.4.4 Disposición espacial 330
5.5.4.5 Información. 331
5.5.4.6 Comportamiento en las crisis 331
5.5.4.7 Recursos. 331
5.5.4.5 Supuestos y tesis 331
5.5.4.5.1 Estatus social. 331
5.5.4.5.2 Tripulación 332
5.5.4.5.3 Proximidad espacial 332
5.5.4.5.4 Transmisión de información 332
5.5.4.5.5 Obstáculos 332
5.5.4.5.6 Tamaño del grupo 333
5.5.4.5.7 Familias 333
5.5.4.5.8 Asignación. 333
5.5.4.6 Conjuntos de datos 333
5.5.4.6.1 Preparación del conjunto de datos 334
5.5.4.6.2 El conjunto de datos y las variables existentes 335
5.5.4.6.3 Creación de nuevas variables 340
5.5.4.6.3.1 Reducción del título 341
5.5.4.6.3.2 Maternidad 341
5.5.4.6.3.3 Niños y ancianos 342
5.5.4.6.3.4 Grupos de edad. 342
5.5.4.6.3.5 Tamaño de la familia y viajes en solitario 343
5.5.4.6.3.6 Tarifa 345
5.5.4.7 Análisis tabulares 346
5.5.4.8 Análisis gráficos. 351
5.5.4.9 Discusión de los análisis exploratorios en el Titanic 357
5.5.5 Liderazgo en contextos educativos 362
5.5.6 Un experimento de variabilidad del ritmo cardíaco 375
5.5.6.1 ¿Qué modelo es el mejor?. 387
5.5.6.2 ¿Qué aprendemos de esto? 387
5.5.6.3 ¿Dónde está la integración del método?. 388
5.5.7 Potenciales de recuperación en la terapia de la adicción 388
5.5.7.1 Ponderación de los acontecimientos vitales. 390
5.5.7.2 Trayectorias de desarrollo. 391
5.5.7.3 Equilibrio entre recursos y vulnerabilidades 392
5.5.7.4 Contribución de los AED. 394
5.5.7.5 Ampliación del AED hacia la estadística de Bayes 395
5.6 Debate sobre el AED. 396

Capítulo 6: *Inferencia Plausible - Estadística Bayesiana*. **399**
6.1 Objetivos del capítulo. 399
6.2 Problemas básicos - Incertidumbre, estimación, decisión 400
6.3 La estadística de Bayes: una introducción 401

6.3.1 Subjetividad: el principal argumento contra Bayes . . . 405
6.3.2 Estudio de caso: otro experimento con té . . . 406
6.3.2.1 Definición del problema . . . 406
6.3.2.2 Supuestos . . . 406
6.3.2.3 Conocimientos previos . . . 407
6.3.2.4 Laplace y la aplicación del teorema de Bayes . . . 410
6.3.2.5 Replicación y actualización . . . 413
6.3.3 Estudio de caso - Diagnóstico médico . . . 415
6.3.3.1 Definición del problema . . . 415
6.3.3.2 Supuestos . . . 415
6.3.3.3 Información empírica . . . 417
6.3.3.4 Aplicación del teorema de Bayes . . . 417
6.3.4 Caso práctico - Fiabilidad de una prueba COVID-19 . . . 422
6.4 El teorema de Bayes . . . 426
6.4.1 Fondo histórico . . . 426
6.4.2 Derivación . . . 427
6.4.2.1 Regla del producto . . . 427
6.4.2.2 Teoría de conjuntos . . . 428
6.4.2.3 Árbol de decisión . . . 430
6.4.2.4 Diagrama . . . 431
6.4.3 Importancia del teorema de Bayes . . . 431
6.5 Excurso sobre la subjetividad . . . 433
6.5.1 Excurso sobre la física: no está muy lejos lo objetivo de lo subjetivo . . . 433
6.5.1.1 Criterio de verdad relativa . . . 435
6.5.1.2 Falta de búsqueda de contrapruebas . . . 435
6.5.1.3 Los modelos no son la realidad . . . 436
6.5.1.4 Integración en un sistema . . . 436
6.5.1.5. Conclusiones de generalización inadmisibles . . . 437
6.5.1.6 Supuestos implícitos sin base empírica de datos . . . 439
6.5.1.7 Como punto (no del todo) final... . . . 440
6.5.1.8 Volviendo a la teoría de Bayes . . . 442
6.5.2 Las creencias subjetivas en la estadística bayesiana . . . 443
6.5.2.1 Creencias subjetivas . . . 445
6.5.2.2 Elección objetiva . . . 445
6.5.2.3 Bayes empírico . . . 446
6.5.2.4 Conclusión intermedia: ¿subjetiva u objetiva? . . . 449
6.6 El vínculo entre el teorema de Bayes y la estadística clásica . . . 451
6.6.1 Comprensión intuitiva de la probabilidad . . . 452
6.7 Posibilidades de los métodos de análisis de datos . . . 462
6.8 Visión general de los métodos del trabajo bayesiano . . . 463
6.8.1 Factores de Bayes y pruebas de hipótesis de Bayes . . . 464
6.8.1.1 Coeficiente de determinación bayesiana R^2 . . . 479
6.8.1.2 Funciones de pérdida . . . 484
6.8.1.3 Caso práctico de función de pérdida . . . 485
6.8.1.4 Actualidad de los factores de Bayes: la seleccion de la Prior . . . 491
6.8.1.5 Ejemplo de investigación: ¿la clarividencia? . . . 504
6.8.1.6 Crítica del diseño - Estudio de Bem . . . 521
6.8.1.7 Factores de Bayes: ¿y ahora? . . . 530
6.8.2 Criterios de información . . . 542
6.8.3 Sobreajuste y subajuste . . . 549
6.8.4 Modelización y estimación por intervalos . . . 556
6.8.4.1 Estimación bayesiana por intervalos . . . 556
6.8.4.2 ROPE – Tolerancia en la estimación puntual . . . 560
6.8.4.2.1 *Caso práctico ROPE - Datos de Darwin* . . . 569
6.8.4.3 Calidad de la predicción y estimación del modelo . . . 574

6.8.4.4 Tasas de aprobación de muestras de casos . . . 578
6.8.4.5 Evaluación gráfica de modelos . . . 584
6.8.4.6 Estudio de caso - Humor y la producción de palabras . . . 599
6.9 Replicación y teorema de Bayes . . . 611
6.10 Deducción frente a inducción . . . 611
6.11 Integración de métodos y métodos mixtos . . . 617
6.12 Elección de las distribuciones a priori . . . 618
6.12.1 La distribución Beta . . . 627
6.12.2 Relación Prior - Likelihood - Posterior . . . 635
6.13 Simulaciones Markov Chain Monte Carlo - MCMC . . . 642
6.13.1 Los algoritmos MCMC . . . 648
6.13.1.1 Algoritmo Metropolis-Hastings . . . 648
6.13.1.2 Muestreo de Gibbs . . . 650
6.13.1.3 HamiltonMonte Carlo . . . 651
6.13.1.3.1 Características de la dinámica hamiltoniana . . . 653
6.13.1.3.2 El paso de leapfrog . . . 653
6.13.1.3.3 La evaluación de la aceptación de Metropolis . . . 654
6.13.1.3.4 La elección de los parámetros de salida y los problemas . . . 655
6.13.1.3.5 Otros desarrollos del HMC . . . 655
6.13.1.4 Resumen de los algoritmos MCMC . . . 656
6.13.2 Ejemplos de algoritmos MCMC en R . . . 659
6.13.2.1 Metropolis-Hastings en R . . . 659
6.13.2.2 Muestreo de Gibbs en R . . . 664
6.13.2.2.1 Muestreo de Gibbs con JAGS . . . 667
6.13.2.2.2 Otros gráficos para el muestreo de Gibbs . . . 672
6.13.2.3 Hamilton Monte Carlo en R . . . 673
6.13.2.3.1 Código del modelo R según Neal (2011) . . . 673
6.13.2.3.2 Estimación de una distribución normal bivariante . . . 676
6.13.2.3.3 HMC comparao con el algoritmo MH . . . 695
6.13.3 Diagnóstico de las cadenas MCMC . . . 700
6.13.3.1 El diagnóstico de Gelman-Rubin . . . 703
6.13.3.2 El diagnóstico de Geeweke . . . 704
6.13.3.3 El diagnóstico de Heidelberg-Welch . . . 705
6.13.3.4 El diagnóstico de Raftery-Lewis . . . 710
6.13.3.5 Análisis mediante gráficos de autocorrelación . . . 711
6.13.3.6 Tamaño efectivo de la muestra . . . 712
6.13.3.7 Resumen de los diagnósticos MCMC . . . 715
6.13.4 Caso práctico: Fisher recargado - más té . . . 715
6.13.4.1 Enfoque frecuentista . . . 716
6.13.4.2 Enfoque bayesiano . . . 717
6.13.4.3 Lady Muriel y MCMC con BUGS . . . 726
6.14 Máxima entropía . . . 728
6.14.1 ¿Qué es la información? . . . 729
6.14.2 ¿Qué es la entropía? . . . 730
6.14.3 Estudio de caso: "Yo, nosotros y la nación" - autoanuncio presidencial . . . 736
6.14.4 Máxima entropía y estadística bayesiana . . . 739
6.14.5 El clásico: ¿es justo un dado? . . . 740
6.14.6 Utilización de información cualitativa para una Prior . . . 743
6.15 Casos practicos bayesianos . . . 746
6.15.1 Las alturas de presidentes . . . 746
6.15.2 Estudio de caso: índices de aprobados en el tratamiento de drogodepencia . . . 771
6.15.2.1 Evaluación a largo plazo 1992-2017 . . . 779
6.15.3 Estudio de caso: "Yo, nosotros y la nación" - autoanuncio presidencial parte 2 . . . 788
6.15.3.1 Estimación MCMC mediante JAGS . . . 788
6.15.3.2 MCMC por fuerza bruta – variante 1 . . . 795

6.15.3.3 MCMC por fuerza bruta – variante 2. 799
6.15.3.4 Integración numérica o aproximación mediante un cuadrículo 802
6.15.3.5 Solución analítica exacta (variante 1) 804
6.15.3.6 Solución analítica exacta (variante 2) 806
6.15.3.7 Factor de Bayes 814
6.15.3.8 Prueba t bayesiana según Bretthorst (1993) 817
6.15.3.9 Resumen 819
6.16 Hacer accesible el conocimiento de expertos sobre la distribución d la Prior. 820
6.17 Conclusión: Lac estadística bayesiana 822

Capítulo 7: *Discusión de las Estadísticas* **825**
7.1 Interpretación de las estadísticas 825
7.1.1 Cosmovisión de la filosofía de la ciencia 825
7.1.2 Diseño de la investigación 826
7.1.3 Muestra y datos 826
7.1.4 Matemáticas puras 826
7.1.5 Estadísticas 826
7.1.6 Conclusiones y causalidad 827
7.2 La elección del enfoque 827
7.3 La pregunta de investigación 828
7.4 Documentación 828
7.5 Salida de datos completos y AED 828
7.6 Renuncia a los rituales y apertura a la flexibilidad 829
7.7 Perspectivas múltiples. 829
7.8 Tamaños de los efectos 830
7.9 Simulación. 830
7.10 Replicación - Replicación - Replicación. 831
7.10.1 ¿Por qué la replicación? 831
7.10.2 Contabilidad de costes totales 832
7.11 Límites de los análisis de datos – la traducción interminable. 833
7.11.1 Orientación al contenido y procesos de traducción811
7.11.2 Garantía de calidad 835
7.11.3 Automatización 835
7.11.4 Reconocer el significado 836

Parte III: Métodos Cualitativos

Capítulo 8: *¿Qué es lo Cualitativo en Realidad?* **841**
8.1 Preludio. 841
8.2 Métodos cualitativos. 842
8.3 Estructura y objetivos de aprendizaje 845

Capítulo 9: *El Paradigma de la Codificación para Analizar Datos Cualitativos* **849**
9.1 Codificación cualitativa y análisis de contenido 849
9.2 El análisis de contenido latente – el análisis cualitativo de textos 851
9.2.1 Reducción - Categorización y codificación. 851
9.2.2 Reconstrucción de sistemas de significado subjetivos 852
9.2.3 Comparación de casos 852
9.2.4 Codificación cualitativa y análisis de contenido 853
9.2.4.1 Enfoque abierto. 853
9.2.4.2 Orientación por la hipótesis 853
9.2.4.3 Enfoque cerrado 854
9.3 La teoría fundamentada (Grounded Theory) 856

9.3.1 Codificación 858
9.3.2 Reducción. 860
9.3.3 Codificación selectiva 862
9.4 Análisis de tablas según Miles y Huberman 863
9.5 Identificación de estructuras, conexiones y tipologías 866
9.5.1 Jerarquías de categorías individuales 867
9.5.2 Secuencias de determinadas categorías 867
9.5.3 Agrupaciones de categorías específicas 868
9.5.4 Búsqueda en el contexto de las categorías asociadas 868
9.5.4.1 Contexto de las categorías individuales relevantes 868
9.5.4.2 Contexto de al menos dos categorías 869
9.5.5 Buscar comprobando determinadas relaciones 869
9.5.5.1 Comprobación de secuencias de codificación simples 869
9.5.5.2 Examinar secuencias de codificación complejas 870
9.5.6 Comparación permanente 871
9.6 Paradigma de codificación de estudio de casos. 871
9.6.1 Cuestionario -- Adicción – Carta de solicitud. 872
9.6.2 Estrategia de análisis con el paradigma de la codificación 873
9.6.3 Codificación con AQUAD 7. 874
9.6.4 Metacódigos 880
9.6.5 Codificación de secuencias 881
9.6.6 Resultados descriptivos 883
9.6.7 Interpretación. 887
9.6.8 Conclusión 888

Capítulo10: *El Análisis del Contenido Manifiesto – El Análisis Cuantitativo de Textos* **889**
10.1 Estudio de caso: Análisis cuantitativo de textos 890
10.1.1 Importatción de textos. 891
10.1.2 Preparación del texto y trabajo preliminar para los análisis 892
10.1.3 Frecuencias de palabras 898
10.1.4 Sensibilidad al contexto: palabras clave contextualizadas 880
10.1.5 Colocaciones 906
10.1.6 Conclusión de la carta de solicitud 910
10.1.7. Ventajas del análisis cuantitativo de textos 911

Capítulo 11: *El Paradigma de la Reconstrucción* **913**
11.1 Prólogo 913
11.2 Hermenéutica objetiva y análisis secuencial 915
11.3 Práctica y práctica vital 916
11.4 Reconstrucción a partir de las huellas 917
11.5 El análisis de secuencias como herramienta central de investigación 919
11.5.1 Análisis secuencial y estructura del caso 920
11.5.2 Control metodológico y falsación 920
11.5.3 El significado latente 921
11.6 El ámbito de la Hermenéutica Objetiva 921
11.7 Concentración metodológica en lo esencial 922
11.8 Un estudio de caso de la práctica terapéutica 923
11.9 Práctica metodológica del análisis de secuencia 926
11.9.1 Generación de hipótesis 927
11.9.1.1 Sobre el fondo teórico. 927
11.9.1.2 Principios de generación de hipótesis. 928
11.9.2 Reglas de interpretación. 928
11.9.2.1 Libertad del contexto 928
11.9.2.2 Literalidad 928

11.9.2.3 Secuencialidad 929
11.9.2.4 Extensividad o totalidad 930
11.9.2.5 Parsimonia 930
11.9.3 Condensación de hipótesis en lecturas del texto 931
11.9.4 Generación de una hipótesis proposicional de estructura de casos. 932
11.9.5 Comprobación crítica de hipótesis 933
11.9.6 Errores típicos en la realización del análisis de secuencias 933
11.9.6.1 Presión interna para terminar 933
11.9.6.2 Libertad contextual 934
11.9.6.3 Literalidad 934
11.9.6.4 Secuencialidad 935
11.9.6.5 Extensividad 935
11.9.6.6 Parsimonia 935
11.10 Hermenéutica objetiva y teoría fundamentada 935
11.11 Integración del métodos. 936
11.12 Análisis de secuencias asistido por ordenador con AQUAD 7 937
11.13 Análisis de secuencia – Ejemplo de un estudio de caso 938
11.13.1 Análisis del contexto 938
11.13.1.1 Situación general de los adictos en abstinencia 938
11.13.1.2 Expectativas típicas de una carta de solicitud 939
11.13.2 La carta de solicitud en detalle 942
11.13.2.1 La unidad de análisis 942
11.13.2.2 Análisis secuencial del texto 942
11.13.2.3 La falsación de la hipótesis preliminar de la estructura del caso. 950
11.13.2.4 Discusión sobre la hipótesis (preliminar) de la estructura de caso 954
11.14 Discusión: Hermenéutica Objetiva 955

Parte IV – Métodos Lógicos

Capítulo12: *Minimización Booleana o Análisis de Implicantes* **959**
12.1 Propedéutica 959
12.2 Ideas básicas de la minimización booleana 962
12.2.1 Causalidad. 963
12.2.2 Estadísticas 963
12.2.3 Principio de exclusión lógica 964
12.3 La formación de tipos como principio de comparación via la minimización lógica 965
12.4 Implicantes primarios y esenciales 972
12.5 Secuencia de pasos de la minimización lógica. 974
12.6 Análisis de criterios - resultado positivo y negativo 975
12.7 Fuzzy logic / Lógica difusa 980
12.8 Valor añadido mediante el análisis de implicantes 982
12.9 Otros parámetros de análisis 984
12.9.1 Variantes de datos 984
12.9.2 Tipos de solucion del QCA. 984
12.9.3 Procedimientos y métodos 985
12.9.3.1 Supuestos simplificados 985
12.9.3.2 Calibración. 985
12.9.3.3 Prueba de necesidad 986
12.9.3.4 Prueba de suficiencia. 986
12.9.3.5 Factorización 986
12.9.3.6 Pruebas estadísticas 987
12.10 Causalidad 987
12.11 Casos prácticos de minimización lógica 988

12.11.1 Sobre la representatividad de las mujeres en los parlamentos 989
12.11.2 Vivir y morir en el Titanic - Parte II 995

Parte V – Síntesis

Capítulo 13: *Combinación de Métodos* **1001**
13.1 En el país de la leche y la miel 1001
13.2 Sobre la complementariedad de los métodos cuantitativos y cualitativos 1004
13.3 Modelos de combinación de métodos 1007
13.3.1 CUAL y CUAN en los diseños de conversión 1007
13.3.2 Formas de conversión de datos 1009
13.3.2.1 Conversión de datos cotidianos 1009
13.3.2.2 Convertir datos cualitativos en datos de frecuencia 1010
13.3.2.3 Conversión de los resultados del análisis cualitativo en datos CUAN 1010
13.4 CUAL y CUAN en diseños secuenciales 1011
13.4.1 Macrosecuencias de la combinación de métodos 1011
13.4.1.1 El modelo de estudio preliminar 1001
13.4.1.2 El modelo de generalización 1011
13.4.1.3 El modelo de profundización 1012
13.4.1.4 El modelo de transformación 1012
13.4.2 Aplicación simultánea de métodos cualitativos y cuantitativos 1013
13.4.2.1 El modelo de triangulación 1013
13.5 Síntesis comparativa de diferentes métodos analíticos 1015
13.5.1 Paradigma de codificación 1015
13.5.2 Análisis cuantitativo de textos 1015
13.5.3 Análisis de secuencia 1017
13.5.4 Comparación de los enfoques 1017
13.6 Esquema de la integración de un método 1018
13.6.1 Pregunta 1018
13.6.2 Preparación y codificación de los datos 1019
13.6.3 Estrategia de análisis de datos y relación con la pregunta de investigación 1020
13.6.4 Integración de datos 1021
13.6.5 Interpretación metodológica 1021

Referencias **1025**

Parte VI – Apéndices (Archivo Separado)

A: El Software de Código Abierto AQUAD 7
A.1 Sumario
A.2 Módulos del programa
A.3 Unidades de análisis de datos yPosibilidades de análisis de datos
A.3.1 Texto, audio, vídeo e imagen
A.3.2 Implementación completa del paradigma de codificación
A.3.3 Implementación completa del análisis de secuencias
A.3.4 Minimización lógica
A.3.5 Análisis combinado de datos cualitativos-cuantitativos y R

B: Programas Informáticos, Guiones y Estudios de Casos: Fuentes en Línea, Descargas y Uso

B.1 Programas estadísticos
B.2 Lista de paquetes R citados
B.3 Información de la sesión de R, scripts de R y su uso
B.4 R-GUI y otros entornos de desarrollo
B.5 Guiones de R
B.5.1 Guiones R de otros autores
B.5.2 Funciones auxiliares generales
B.5.3 Parte I Filosofía de la Ciencia
B.5.4 Parte II Estadísticas clásicas
B.5.5 Parte II Estadísticas exploratorias
B.5.6 Parte II Estadística Bayesiana
B.5.7 Parte III Métodos cualitativos
B.5.8 Parte IV Métodos lógicos
B.5.9 Síntesis de la Parte V

C: Estudios de Casos y Conjuntos de Datos

C.1 Estadísticas clásicas
C.2 Análisis exploratorio de datos
C.3 Estadística bayesiana
C.4 Paradigma de codificación
C.5 Análisis cuantitativo de textos
C.6 Análisis de secuencias
C.7 Minimización booleana

Prólogos

»Shake your rump-a.«

Album »Paul's Boutique«
Beastie Boys, 1978–2012

Los Beastie Boys éran un grupo de hip-hop cuya música encaja perfectamente con el tema de este libro. Originalmente comenzaron a finales de los años 70 del siglo XX como una banda de HC punk, haciendo hip-hop y a lo largo de los años integrando una gran variedad de estilos de música como el rock alternativo e independiente, el jazz, el funk, el punk rock, el electro, latín, cantos tibetanos y muchos más, por no hablar de la gran cantidad de muestras de otros músicos y los músicos, que fluyen en sus canciones. Mientras tanto, lanzaron otro álbum de punk o temas sueltos al estilo del punk rock, para volver a orientarse hacia el hip hop y la música experimental. ¿De qué tipo de música queremos hablar? ¿Qué nombre queremos dar al estilo de los Beastie Boys? ¿Ha evolucionado el grupo, ha cambiado, se ha ampliado? ¿Qué significa combinar diferentes estilos de música, mezclar muestras e integrarlos en algo nuevo, de modo que no quieras usar realmente los acompañamientos estilísticos de dónde provienen las partes individuales? Los Beastie Boys no eran ciertamente lo único grupo que amplió su estilo musical con el tiempo, utilizó samples, creó mezclas, etc. - y aún así se mantuvo la coherencia. Más tarde, en la década de 2000, llegaron los mash-ups, que mezclaron canciones enteras y no sólo usaron samples. Podemos aprender mucho de todos estos grupos y DJs para el tema de nuestro libro: Mixed Methods y la combinación de métodos (de análisis de datos).

Los métodos mixtos y la triangulación forman parte del repertorio estándar de las ciencias sociales desde hace muchos años. Han entrado de lleno en la corriente principal. Esto llega hasta el punto de que uno tiene la impresión de que ahora es prácticamente parte del buen tono en muchos proyectos de calificación, como si esto por sí solo mejorara la calidad de la obra o produjera algo que de otro modo nunca habría salido a la luz. Esto no es culpa de los estudiantes, sino, sobre todo, a las nuevas exigencias y expectativas de la comunidad científica. En realidad, consideramos que tal desarrollo es inapropiado y se pierde el punto. Porque los métodos no deben estar por encima de las preguntas. Cuando esto sucede, algo ha ido muy mal durante mucho tiempo, a menos que se examinen los propios métodos. Entonces, por supuesto, pertenecen al número uno.

¿Cuándo y por qué exactamente los métodos van juntos? Desde nuestro punto de vista, una respuesta adecuada sería que si la cuestión lo sugiere y lo requiere con urgencia, nada se opone a una combinación de métodos. De lo contrario, es mejor no hacerlo. Investigar bien una cuestión de forma empírica con un solo método o procedimiento es todo un reto y puede conducir a hallazgos significativos, como las curvas de memoria de Hermann Ebbinghaus o los experimentos con palomas de Burrhus F. Skinner. No había nada que combinar, la investigación era de primera clase y los resultados siguen estando entre los fenómenos realmente bien estudiados y, en general, entre los fundamentos del procesamiento de la percepción y el comporta-miento. E incluso pueden replicarse, lo que es más de lo que puede decirse de muchos otros estudios.

Pero imaginemos que una pregunta obliga ahora prácticamente a una sabia combinación de métodos: consigo mismo, con otros, con el tiempo, con las personas, con los métodos, etc. ¿Qué se necesita ahora? Consideramos esencial un conocimiento profundo que incluya habilidades prácticas de todos los métodos

de recogida y análisis de datos que se utilizan actualmente. No sólo hay que entenderlos, dominarlos en la práctica y adaptarlos con flexibilidad a las necesidades, sino que también hay que comprender el respectivo trasfondo epistemológico. De lo contrario, se combinan métodos desde una sola perspectiva epistemológica sin ni siquiera notar esta orientación unilateral. No se trataría entonces ni de una integración ni de una combinación, sino simplemente de la interpretación de diferentes métodos sobre la base de un único paradigma epistemológico. La integración real se parece más a un proceso dialéctico o incluso tetralémico de resolución de problemas, en el que al principio no está nada claro lo que saldrá después, pero, por favor, de forma constante y replicable y bien fundamentada. A cambio, uno se mueve durante mucho tiempo con la máxima incertidumbre, lo que requiere mucha creatividad y capacidad de improvisación, además de resistencia. El lado positivo es que más tarde se obtienen nuevos conocimientos que hacen olvidar el arduo viaje. Un ejemplo de nuestra propia práctica de investigación: nosotros mismos realizamos las entrevistas en un proyecto de investigación anterior, las transcribimos, las codificamos, las ordenamos y las categorizamos, comprobamos las hipótesis sobre el texto, etc. Además, hubo análisis lógicos de minimización booleana y algunos análisis de tablas numéricas. A pesar de todos estos esfuerzos, se necesitó casi un año de trabajo extenuante sobre el material en el proceso de investigación hasta que surgió una estructura a partir del material textual, lo que en realidad condujo

a una tipología justificable. Incluso cabe en un pequeño trozo de papel y todavía hoy parece coherente. El tiempo y la energía fueron bien invertido. Sin embargo, el número de categorías a utilizar superaba no sólo nuestra memoria a corto plazo, sino que también supuso una enorme carga para nuestra memoria a largo plazo. Sin ordenador no habríamos podido hacerlo en absoluto. Al mismo tiempo, rompimos algunas reglas que proponemos seriamente en este libro: por ejemplo, limitar la cantidad de códigos y categorías. Pero aún más difícil es el tiempo que se pasa trabajando y al mismo tiempo esperando, sin poder tener el objetivo o incluso la solución exactamente en mente.

La hermenéutica objetiva dice que en una crisis, las rutinas fallan y las estructuras de casos individuales, el propio habitus, se hacen especialmente evidentes. Vemos la investigación como una crisis en muchos aspectos, aparte de que no sabemos qué saldrá si realmente queremos investigar y aplicar algo nuevo. En cambio, ¡déjate sorprender! Para ello, tenemos que cambiar y hacer las cosas de una manera nueva o, al menos, diferente, sin negar inmediatamente o incluso negar lo antiguo.

Con todo esto, hemos escrito el libro para abordar las crisis a las que uno debe enfrentarse, especialmente cuando se trabaja con la estadística clásica o bayesiana, el análisis de datos exploratorio, cualitativo o lógico. No nos limitamos a abordar el pensamiento subyacente, las opiniones generales y a veces completamente contradictorias de otros investigadores. Más bien tratamos de arrojar algunas piedras en el camino de los lectores. En las artes marciales asiáticas, por ejemplo, se acostumbra no sólo a arrojar figuradamente piedras en el camino de los alumnos para que aprendan a recogerlas y seguir adelante sin prestar más atención a las piedras de la necesaria.

Por tanto, nuestra obra no es un libro de texto. Está pensada para servir de inspiración y quizás para ampliar la propia autonomía, para tomar y representar decisiones independientes de las modas. Por eso siempre intentamos nombrar los antecedentes de los procedimientos y dar fórmulas en algunos casos. Para los contenidos en detalle, existen excelentes libros de texto en número y variedad suficientes. Nos limitamos a las áreas de los procedimientos que son problemáticas o particularmente ventajosas desde nuestro punto de vista. Dejamos de lado los detalles más profundos, ya que a menudo sólo se exploran en artículos de revistas muy detallados. Y para todo no hay realmente una respuesta o solución general y siempre válida. Por ello, exploramos algunos conjuntos de datos para clarificarlos utilizando diferentes métodos de análisis de datos para mostrar precisamente eso: el mundo y nuestros hallazgos son relativos, no absolutos. Y no sólo obtenemos a veces un resultado muy similar, sino que a veces incluso cambia la pregunta y los resultados son difíciles de comparar. Pero esa es la realidad de la investigación. Desde nuestro punto de vista, no existe un método óptimo para analizar los datos. Los lectores notarán ciertas redundancias, repeticiones y, en algunos lugares, una longitud desagradable en este libro. Pero antes de que se apresure a tachar todo esto de mal estilo, negligencia, ineficacia o incluso incompetencia literaria, están entre las piedras mencionadas. Además, hemos utilizado el método de pensar en voz alta en algunas digresiones, como el estudio de caso sobre el análisis de secuencias, para ilustrar los procesos de razonamiento que inevitablemente se producen en el análisis de datos. Nosotros mismos echamos de menos esta figura en muchos libros de texto, que a menudo parecen demasiado lisos y sencillos y, por desgracia, no enumeran las múltiples consideraciones
y los laboriosos pasos individuales uno por uno. Pero esa es precisamente la cuestión: los problemas y las dificultades para mantener la coherencia y la ciencia. Re-aprender es mucho más difícil que aprender algo nuevo. Hay que gastar mucha energía para sustituir lo viejo por lo nuevo, y en el proceso siempre se vuelve a caer en los viejos hábitos, a menudo sin darse cuenta. Si le preguntan al respecto, puede incluso reaccionar de forma poco amable y no sentirse emocionalmente feliz por tales insinuaciones. Por ello, no nos cansamos de mencionar los problemas de la estadística clásica, por ejemplo, siempre que parece necesario. Sin embargo, también hacemos hincapié con la misma vehemencia en los problemas de la estadística de Bayes, como los que pueden surgir del uso irreflexivo de los factores de Bayes o sin una previa anclada en el contenido. Precisamente porque muchos libros de texto de ciencias sociales ni siquiera se ocupan de estos problemas, sino que se limitan a ilustrar procedimientos algorítmicos en la tradición de los libros de Ronald A. Fisher, consideramos legítimo permitirnos algunas redundancias e inconvenientes aquí y allá. Los lectores pueden perdonarnos por esto. Demasiados libros de texto se reprimen noblemente cuando se trata de los problemas del enfoque que ellos mismos defienden -es decir, el tema del libro de texto en cuestión-, a pesar de que todos los argumentos pertinentes pueden leerse con la máxima precisión y detalle en muchos artículos de revistas. Intentamos adoptar una postura cuando es posible, y cuando tiene sentido que lo hagamos. Esto no significa que una opinión diferente no sea legítima. Y, por supuesto, como todas las personas, tenemos nuestras preferencias en cuanto a determinados enfoques metodológicos. Günter Huber, por ejemplo, es un gran fan del paradigma de la codificación y la minimización booleana, y Leo Gürtler siempre disfruta con el análisis de secuencias de la hermenéutica objetiva -sin estadísticas- y la estadística bayesiana. Pero eso no es razón para no trabajar con otros métodos y aprenderlos cuando una pregunta simplemente lo requiera. Así que si las redundancias te molestan, puedes simplemente omitirlas o que te molesten. A nosotros, en cambio, no nos molesta en absoluto.

17
Leo Gürtler & Günter L. Huber
(Karlsruhe & Tübingen, 2006–2022)

Agradecimientos

»Bodhisattas complete the danaparami or perfection of giving to the ultimate degree by happily donating their limbs and their very lives to help other beings.«

Dana—The Practice of Giving
Susan Elbaum Jootla, 1945–

Cada libro tiene una historia y un trasfondo sin los cuales nunca habría sido escrito. Los libros y la ciencia demuestran de forma impresionante que sólo se puede conseguir algo juntos, aunque parezca diferente en la superficie y normalmente sólo haya unos pocos autores en el título del libro. Sin los innumerables artículos científicos y fuentes secundarias, sin el código, los tutoriales, los blogs y las discusiones en los foros, sin la comunicación y las consultas con la gente y sin un entorno informático que funcione o programas para crear la maquetación, las entradas bibliográficas, etc., es imposible crear un libro sensato y serio. Y todas las fuentes mencionadas tienen a su vez sus propias fuentes en las que se basan, etc. En nuestro caso, sin el trabajo de todas las personas que de alguna manera transmiten su trabajo a los demás, no habría sido posible ni siquiera una fracción del libro. Esto es bastante independiente de si hablamos de conocimiento científico, código, soporte informático, etc. o de otras influencias que

han conformado nuestra visión de los métodos mixtos o lógicos, cualitativos y cuantitativos. Damos las gracias a todas estas personas.

Dado que un libro como éste supone un gran esfuerzo durante años, no puede hacerse sin una familia que simpatice con este trabajo. Esta influencia no puede ser sobrestimada. Agradecemos a nuestras familias que nos hayan permitido escribir este libro. Luego están nuestras mascotas a lo largo de los años (perros, gatos) que son una fuente de gran alegría. Sin ellos, la vida sería definitivamente más aburrida. No queremos olvidarnos de todas aquellas personas que nos han apoyado mucho académicamente. Hay bastantes y enumeramos las personas que son muy relevantes para nosotros.

Agradecimientos especiales de Günter Huber: El manual del paquete de software Qualog para grandes ordenadores de Anne Shelly y Ernest Sibert, que me regalaron los autores hace más de 30 años, fue la inspiración y el modelo para AQUAD, el paquete de programas para el análisis de datos cualitativos. En aquella época, sólo se daba cuenta de las posibilidades de la llamada "programación lógica" para comprobar los vínculos entre codificaciones. A propuesta de Michael Huberman, se incorporaron posteriormente diversas formas de análisis de tablas y el enfoque de análisis comparativo cualitativo de Ragin (1987). El programa debe numerosas mejoras y ampliaciones, incluida la reprogramación para una versión multi-plataforma, a la colaboración de colegas de Letonia y España (en adelante, por orden alfabético), a sus publicaciones, a las invitaciones a conferencias y talleres en sus universidades y a las interesadas contribuciones de sus estudiantes: Samuel Gento (Madrid), Jordi López (Barcelona), Carlos Marcelo (Sevilla), María-Angeles Martínez (Alicante), Irina Maslo (Riga), Antonio Medina (Madrid), Ramón Pérez (Oviedo), Manuel Saavedra (Morelia/México), Jorge Schmitt (Buenos Aires), Luis Miguel Villar (Sevilla), Miguel Angel Zabalza (Santiago de Compostela). No se olvidan los comentarios y sugerencias de otros numerosos colegas en las conferencias del Centro de Psicología Cualitativa. Mi más sincero agradecimiento a todos ellos.

Leo Gürtler agradece especialmente a las siguientes personas: Urban M. Studer me hizo conocer la estadística de Bayes en primer lugar. También me enseñó los fundamentos del pensamiento bayesiano y, junto con Gerhard Scholz (start again, m&o, musivo), me introdujo en el trabajo con el análisis de secuencias y en la Hermenéutica Objetiva. Robert S. Hankin, el autor del paquete de R Brobdingnag, escribió rutinas para manejar números muy grandes, que resultaron esenciales para implementar el script original de Mathematica R de Urban Studer para el artículo de G. L. Bretthorst. De lo contrario, el script R asociado sólo funcionaría para números pequeños, lo que sería muy desafortunado. El físico del plasma Udo v. Toussaint de IPP/ MPG Munich/ Garching fue muy generoso y nos dedicó una mañana de su tiempo y respondió a muchas preguntas sobre la estadística de Bayes. Esto fue muy inspirador y se quedó con nosotros durante años. Günter L. Huber programó muchos complementos adicionales en AQUAD para que fuera posible investigar de forma semiautomática y sistemática las hipótesis en un gran número de datos de vídeo para realizar análisis cuantitativos adicionales, lo que otros programas de QDA nunca habrían hecho posible. De este modo, también se pueden utilizar métodos mixtos. Además, hay un número incontable de personas de muy diversos ámbitos de la sociedad que han contribuido con sus declaraciones, publicaciones, etc. a mi desarrollo posterior.

Estructura del libro

En realidad, se trata de varios libros sobre métodos de análisis de datos, sus fundamentos, así como las aplicaciones resultantes y sus posibles combinaciones. Por ello, el libro está dividido en varias partes para aumentar la legibilidad y presentar los enfoques por separado. Está estructurado en gran medida de forma que las partes sólo dependen unas de otras de forma limitada y pueden leerse fácilmente por separado. La aplicación práctica de la estadística y los métodos lógicos se realiza con el software de código abierto R (www.r-project.org), utilizando casos prácticos y scripts que examinan los temas tratados estadísticamente o con la ayuda del álgebra de Boole. La parte práctica cualitativa puede seguirse con AQUAD7 o 8 (www.aquad.de). AQUAD7/8 trabaja junto con R para ciertos análisis exploratorios, en su mayoría descriptivos, y pasa scripts a R en segundo plano. Todos los guiones que escribimos están disponibles en YYY. Los conjuntos de datos asociados también pueden descargarse allí, a menos que los derechos de autor lo prohíban a un tercero. Al principio de los estudios de caso y de las demostraciones prácticas con R, siempre se indica entre paréntesis el nombre del script R correspondiente (R-Skriptname.r), que suele ser algo más extenso que el código R impreso. En algunos casos nos remitimos a los scripts de R de otros investigadores. Nuestros scripts están sujetos a la licencia GNU GPL v.3 (para más detalles, véase GNU GPL v.3). Esto significa que son de código abierto y pueden ser utilizados por cualquier persona de forma gratuita. Los guiones y el software de otros autores están sujetos a las respectivas licencias de estas personas y pueden aplicarse a lo largo de estas licencias.

Los capítulos de un vistazo

La parte I comienza con la epistemología y recapitula sus fundamentos.

El capítulo 1 sobre el conocimiento y la sabiduría destaca el carácter relativo de un mundo cambiante y transitorio. Esta relatividad no puede ser anulado por la ciencia. Al mismo tiempo, la cognición y la sabiduría también dependen de la cultura, ya que hay otras formas de cognición en otras tradiciones de vida y pensamiento, que se ponen claramente sobre los criterios científicos externos, es decir, ajenos al ser humano, que hoy en día son generalmente aceptados y que se consideran como la norma. Estos incluyen, entre otros, el conocimiento a través de la experiencia directa aparte del pensamiento en la enseñanza según el Buda.

El capítulo 2 sobre lógica e inferencia se centra en las operaciones de pensamiento básicas del trabajo científico. La inducción, la deducción y la abducción se manejan en su mayoría como operaciones separadas. La tesis es que esto sólo es así a efectos analíticos, pero en la práctica de la investigación estas tres operaciones se fusionan de forma fluida y constante. Pero ahí no acaba la historia, porque la cognición no requiere necesariamente un edificio científico. Podemos simplemente jugar o hacer algo y así aprender y reconocer sea de paso - ¡y no poco!

El capítulo 3 presenta a los que consideramos los filósofos de la ciencia más influyentes: David Hume cuestionó la inferencia por inducción, Popper señaló la necesidad de la falsación y Lakatos amplió el enfoque científico de las teorías dividiéndolas en núcleo teórico central y periferia para proteger la construcción cuidadosa de teorías prometedoras y no descartarlas precipitadamente. Kuhn y Feyerabend ampliaron el enfoque de la construcción interna de las teorías a los aspectos sociales, por ejemplo la cuestión de Kuhn de cómo se alternan los paradigmas o que, según Feyerabend, la ciencia y el proceso científico no son las únicas formas legítimas de conocimiento humano.

La parte II examina los métodos cuantitativos. ¿Qué hacen las estadísticas, qué formas adoptan y cuáles son los problemas en cada caso?

El capítulo 4 presenta la estadística clásica según Fisher y Neyman-Pearson. El punto central es la tesis de que ambos enfoques son incompatibles entre sí, tal y como se enseñan hoy en día y, en la mayoría de los casos, de forma irreflexiva. Fisher quería contribuir al descubrimiento de la novedad y, por tanto, al conocimiento inductivo, mientras que Neymean-Pearson aspiraba al comportamiento inductivo. Las distintas secciones están dedicadas a los elementos fundamentales, como el concepto de probabilidad basado en las frecuencias . Otros temas de debate son la prueba estadística, el concepto de significación y los elementos centrales de las respectivas teorías estadísticas, con el fin de averiguar su significado y propósito y, al mismo tiempo, cuestionar si su uso tiene sentido o no, y cuándo. Estos incluyen el valor p y los umbrales críticos de significación y las direcciones de las hipótesis, el tamaño de la muestra, el tamaño del efecto, los índices de error estadístico y la potencia. Nuestra preocupación es demostrar que el concepto de significación debe abandonarse por completo. Estimar en lugar de probar debería ser el lema. Las simulaciones, por su parte, son una poderosa y excelente herramienta para entender mejor la estadística y poder investigar problemas ambiguos incluso sin una solución analítica. Los conceptos asociados, los fenómenos y las fuentes comunes de error en estadística – replicación, aleatoriedad, sesgos y falacias múltiples, y paradojas- forman parte del repertorio, junto con los datos ausentes, la cuestión de las condiciones previas de los procedimientos y el problema del establecimiento de la equivalencia, cuya discusión acerca a la estadística para poder tratar mejor los problemas reales de la investigación. Cuando es posible, el código R apoya la presentación de los temas.

El capítulo 5 sobre la estadística exploratoria de Tukey muestra que no es necesario buscar la significación para descubrir relaciones y diferencias significativas en los datos, sino que la creatividad y la adecuación de los casos, así como las transformaciones de los datos, son suficientes para derivar hipótesis sustanciales para una investigación posterior en combinación con diferentes fuentes de información. El trabajo estadístico exploratorio se demuestra concretamente mediante varios estudios de casos de la práctica. El principio rector es siempre la idea de examinar los datos por su contenido real en lugar de limitarse a probarlos para confirmarlos, con el fin de encontrar estructuras y patrones. Aunque se utiliza una gran variedad de métodos, todos los estudios de casos tienen en común que prescinden por completo de la estadística inferencial y siguen produciendo conclusiones razonables y significativas.

El capítulo 6 sobre la estadística de Bayes muestra un planteamiento básicamente sencillo en forma de teorema de Bayes, que rápidamente se convierte en una complejidad matemática, por lo que hay que realizar complejas simulaciones para estimar estos modelos, ya que sin el uso del ordenador esto no es posible en absoluto. A diferencia de la estadística clásica, un concepto de probabilidad intuitivamente comprensible determina lo que ocurre y no se trata de significados, sino del comportamiento de las probabilidades en diferentes condiciones. El equivalente bayesiano de la prueba de significación, los factores de Bayes, combinan en gran medida la estadística bayesiana y la clásica y, por tanto, reintroducen directamente muchos problemas de la estadística clásica en la estadística bayesiana por la puerta de atrás. Entre ellas se encuentran el énfasis excesivo en las características matemáticas de las distribuciones en lugar de justificarlas en términos de contenido, la concentración casi fetichista en los valores mínimos de los factores de Bayes para separar la significación de la irrelevancia independientemente de las consideraciones de contenido y, en última instancia, la pérdida del enfoque bayesiano completo. El uso exclusivo de los factores de Bayes está lejos de ser una aplicación completa del enfoque de Bayes. Detrás de estos esfuerzos está, al parecer, el afán humano por los análisis (semi)-automáticos de datos, contra el que advierte expresamente, entre otros, el psicólogo Gigerenzer. Sin embargo, también es posible sin los factores de Bayes, o incluso pueden utilizarse de forma inteligente. Los métodos de trabajo bayesianos, por ejemplo, abarcan el ámbito de la estadística de Bayes. Esto incluye la prueba de hipótesis con la ayuda de la distribución posterior - aquí los factores de Bayes pueden desempeñar un papel razonable -, las estimaciones de intervalos y la estimación de los límites de tolerancia con la ayuda del concepto de ROPE, varios criterios de información, así como soluciones a los problemas de sobreajuste o infraajuste. Un aspecto central de Bayes es el uso de distribuciones a priori, que representan el conocimiento previo a la obtención de los datos. Aquí es donde la estadística se encuentra de paso con los métodos mixtos, ya que no siempre se dispone de datos numéricos de estudios anteriores para formar un previo. A continuación, es necesario realizar traducciones cuidadosas para traducir con precisión el contenido en distribuciones matemáticas a priori. Del mismo modo, las consideraciones de un previo pueden servir para evaluar los estudios en cuanto a su adecuación con

respecto a sus consideraciones teóricas y operacionalizaciones, lo que se ensaya en un estudio de caso sobre el estudio experimental del psicólogo social Bem sobre la clarividencia. La simulación Markov Chain MonteCarlo (MCMC) es la única solución para estimar modelos complejos, ya que no se pueden encontrar soluciones analíticas. Se presentan y discuten los algoritmos MCMC más comunes y se prueba su implementación en R. En varios excursos prácticos y teóricos, investigamos las acusaciones contra la estadística de Bayes, como la subjetividad, y comprobamos su validez. Una característica especial de Bayes es la capacidad de aprender del pasado, lo que convierte a la estadística de Bayes en un enfoque en el que el aprendizaje de la experiencia no sólo es posible, sino que forma parte integral. Aprender de la experiencia demuestra un estudio de caso con R. El enfoque de máxima entropía, por otra parte, es un caso especial en el contexto de Bayes, ya que permite transformar cualquier previa dada según la máxima entropía para crear una previa óptima según los principios de la mecánica estadística o de la termodinámica.Varios estudios de casos que incluyen el análisis de los mismos datos utilizando diferentes algoritmos (solución exacta, simulación MCMC) completan el capítulo.

El capítulo 7 resume las explicaciones sobre las estadísticas. Partiendo de las interpretaciones de las estadísticas, nos preguntamos qué es lo relevante. Para ello, el capítulo aborda repetida y selectivamente los planteamientos ya presentados. Un punto realmente significativo parece residir en la replicación y el cálculo de costes completo, así como en la necesidad de no olvidar el contenido en todos los coeficientes y parcelas. Las réplicas no son tan fáciles de realizar como parece y el cálculo de los costes totales constituye un aspecto que todavía se descuida mucho en la ciencia, que debería añadirse a las características de calidad, pero sin exagerar. El capítulo es subjetivo por nuestra parte, ya que, según la perspectiva de cada uno, la interpretación de las estadísticas es muy diferente. El capítulo concluye con los límites del análisis de datos, es decir, con la tesis de que el trabajo estadístico es una traducción constante de lo cualitativo a lo cuantitativo y viceversa. Los análisis (semi) automatizados deben entenderse como una advertencia, ya que no se deben realizar análisis de datos sin sentido común; y en cuanto se pierde de vista el contenido debido a los coeficientes abstractos, se está en camino de perder lo esencial. Reconocemos el sentido al cuestionar críticamente las convenciones, independientemente del paradigma del que provengan. Para ello, es necesario dedicar tiempo a considerar el significado concreto de un resultado estadístico en la práctica, una tarea no trivial.

La Parte III pasa a los métodos cualitativos, aquellos que trabajan principalmente con información categórica en lugar de numérica, al tiempo que incorporan interpretaciones para formar categorías en primer lugar como base de comparación para la prueba de hipótesis y la inferencia.

El capítulo 8 se adentra en el tema del análisis de datos cualitativos y se pregunta qué constituye la calidad o qué es el elemento cualitativo en el análisis de datos cualitativos. Una pregunta justa, después de todo, en otro lugar donde afirmamos que la estadística consiste en una secuencia casi interminable de traducciones recíprocas de elementos cuantitativos y cualitativos. Y en otro lugar afirmamos que a los investigadores cualitativos les gusta presentar sus resultados de una forma que se asemeja al análisis de frecuencias numéricas en lugar de a la interpretación cualitativa, como cabría esperar en realidad.

El capítulo 9 introduce el paradigma de la codificación, que es más o menos el paradigma dominante del análisis de datos cualitativos, ya que, independientemente del enfoque, todos acaban por llegar al punto en el que necesitan convertir el texto en interpretaciones. A partir de ahí, se puede hablar de la formación de categorías (= codificaciones) y el proceso puede complicarse a voluntad. La codificación puede ser de contenido, es decir, de lo que trata un texto, o de estructura, es decir, de cómo se presenta o comunica algo, por ejemplo. Se puede codificar texto, imágenes, audio y vídeo, lo que tiene en cuenta todas las fuentes de información imaginables. Los códigos pueden combinarse en otros más abstractos o buscarse en cualquier secuencia complicada. La búsqueda de códigos permite comprobar las hipótesis (de secuencia) directamente en el texto y, por lo tanto, ofrece la posibilidad de formular un marco de interpretación con base empírica en diferentes fuentes de datos. De este modo, se pueden formar posteriormente tipologías u otros constructos. La posibilidad de contar los códigos, tabularlos y hacerlos accesibles a cualquier análisis numérico, gráfico y lógico proporciona al paradigma de la codificación

un acceso casi inherente a los métodos mixtos y lo hace de forma bien justificada, es decir, por necesidad. Consideramos que esto es una característica, no un error.

El capítulo 10 representa la parte numérica del análisis de datos cualitativos en forma de análisis de texto cuantitativo. Aunque se podría pensar que esto debería situarse en la estadística, en realidad ya es un tipo de métodos mixtos cuando se aplica con sabiduría y sensibilidad al contexto. La preparación del texto es un factor crucial, ya que se trata de lo que podría eliminarse del texto para facilitar la búsqueda de diferencias y similitudes. El análisis cuantitativo de textos ofrece todo el repertorio del análisis numérico, pero depende de lo bien que se formule y prepare el punto de partida del análisis. Así, se pueden incluir palabras clave, palabras basadas en el diccionario, frases y combinaciones de palabras con y sin terminaciones, o la búsqueda de colocaciones y la cuestión de la proximidad y la distancia de los conceptos interesantes. Si estos análisis se combinan con la idea básica del paradigma de la codificación, a saber, transferir el texto a objetos de interpretación (= codificaciones, categorías), incluso enormes cantidades de datos pueden analizarse de forma muy eficiente preservando su complejidad, por ejemplo, de los medios sociales.

El capítulo 11 se centra en la reconstrucción del significado latente según la metodología de la hermenéutica objetiva y, especialmente, con el procedimiento del análisis de secuencias. Este método puede considerarse como la contrapartida cualitativa de la estadística bayesiana: se arregla con muy pocos datos y siempre funciona. En un primer paso, el análisis secuencial elabora una estructura de sentido latente sobre el caso de forma metodológicamente controlada y la comprueba de nuevo sobre el texto en un segundo paso para llevar a cabo la necesaria falsación de la llamada hipótesis de estructura de caso preliminar. Con un buen trabajo preliminar, los pasajes críticos del texto apoyan esta estructura elaborada y no la falsean. Como resultado, el investigador recibe una estructura de casos probada. El estudio de caso, una carta de solicitud de baja en psiquiatría para una plaza de terapia de adicción en régimen de internado, trabaja con el peor de los casos, un texto muy breve, y además analiza sólo fracciones del mismo. El objetivo es mostrar lo poderoso que es este método y lo bien que funciona a pesar de las circunstancias adversas. El método de pensar en voz alta se utiliza para llevar a cabo los distintos pasos del análisis. El estudio de casos también se utiliza para el paradigma de la codificación y el análisis cuantitativo de textos, de modo que puedan establecerse comparaciones entre los procedimientos y las conclusiones resultantes.

La parte IV complementa el mundo de la información numérica y cualitativa con el de la lógica. Los análisis no se basan en un más o menos, ni en la pura interpretación, sino en la aplicación exclusiva de principios y operaciones lógicas.

El capítulo 12 introduce la minimización booleana, también llamada análisis de implicantes. Permite elaborar configuraciones condicionales para formular configuraciones condicionales típicas para la ocurrencia (o no) de un criterio de interés a través de los casos o también en el contexto de los meta-análisis. Es importante analizar siempre los datos según el criterio positivo y el negativo (invertido) y comparar los resultados para que no surjan soluciones contradictorias que pongan en duda las interpretaciones.

La parte V está dedicada a los métodos mixtos y a la cuestión de cómo podría ser una combinación exitosa de métodos.

El capítulo 13 se plantea si, cuando se combinan los métodos, se vive realmente en una tierra de leche y miel o se afronta un gran reto que quizá no se pueda resolver en absoluto. Combinar métodos no significa simplemente tomar una fuente de información aquí y un método de análisis de datos allá y agruparlos. Como debería haber quedado claro en el transcurso del libro, hay una perspectiva epistemológica más o menos diferente detrás de cada método y cada análisis, que no puede combinarse sin más con otro. ¿Pueden combinarse en absoluto y qué posibilidades de una secuencia combinada de clases de métodos existen en la literatura? ¿Hay que ver los métodos por separado o el elemento combinatorio es quizá ya

inherente a ellos? Así, podría decirse que CUAN y CUAL son complementarios y, como toda buena figura tetralémica o dialéctica, se trata de integrar la información en una nueva perspectiva superior, que puede no responder a la cuestión de la perspectiva epistémica, pero sí marcarla como irrelevante a partir de entonces. Por lo tanto, como conclusión, el estudio de caso de la carta de solicitud de adicción -analizada con el paradigma de codificación, el análisis cuantitativo de textos y el análisis de secuencias- se utiliza como ejemplo de comparación. Nos preguntamos cuál de los métodos aplicados llega a qué resultados y cómo pueden conciliarse entre sí. A continuación, se analizan simultáneamente las ventajas y desventajas de los métodos analizados. El capítulo concluye con un esbozo metodológico de cómo el estudio de caso cualitativo-cuantitativo de la estadística bayesiana puede ampliarse para analizar el uso del lenguaje en los duelos americanos pre-soviéticos. Esto implica la pregunta de investigación, la estrategia, la integración de los datos y la interpretación.

La bibliografía comienza en la p. 1025, enumera la literatura citada en el texto.

La parte VI (archivo separado) examina los materiales concretos y las herramientas auxiliares que se presentan y aplican en el libro. El contenido de La parte VI es resumido en un archivo separado, ya que se trata de apéndices.

El apéndice A presenta el software de código abierto AQUAD 7, escrito por uno de los autores (Günter Huber) y que tiene al otro autor (Gürtler) como proveedor de ideas y diseñador. Además de los módulos del programa, se discute la implementación, es decir, el paradigma de codificación, el análisis de la secuencia, el análisis de los implicados y los vínculos selectivos con R.

El apéndice B enumera los recursos utilizados. Esto incluye el software utilizado (R, JAGS, BUGS y Stan) y toda la información necesaria para comprender la versión y el entorno de software de los programas utilizados. El objetivo es que todos los aspectos relacionados con el software del libro puedan reproducirse con exactitud. A continuación, como conclusión del libro, se enumeran todas las fuentes de información y datos que hemos utilizado para los estudios de caso detallados en los respectivos capítulos.

El apéndice C ofrece una visión general de los estudios de casos y conjuntos de datos utilizados en el libro, que proceden de una amplia variedad de fuentes, estudios empíricos y conjuntos de datos ya disponibles en la R.

Parte I

Comprensión y Sabiduría

Capítulo 1

Discurso Precursor de la Teoría de Ciencia

"La dialéctica es la formación del espíritu de contradicción, que se da al hombre para que aprenda a reconocer la diferencia de las cosas"

(Die Dialektik ist die Ausbildung des Widersprechungsgeistes, welcher dem Menschen gegeben, damit er den Unterschied der Dinge erkennen lerne.)

Über Natur und Naturwissenschaft.
Johann Wolfang von Goethe, 1749–1832

¿Qué necesitamos para poder trabajar científicamente? Según la cita anterior de Goethe, necesitamos opuestos y diferencias. Si se le preguntara a alguien hoy, la respuesta podría ser que necesitamos información. Goethe era un espíritu muy progresista, porque según Bateson (1985), la información no es otra cosa que las diferencias que marcan las diferencias y así adquieren sentido. Las contradicciones, en cambio, dan lugar a la resolución de las mismas, lo que es posible a través del trabajo dialéctico. Para ello es necesario integrar la información obviamente contradictoria desde una perspectiva diferente, de modo que ya no se contradigan entre sí, sino que puedan conciliarse para formar un modelo coherente y consistente.

Ahora podríamos preguntarnos: ¿conforme a quién queremos trabajar científicamente y si queremos "trabajar conforme a alguien"? ¿Es Goethe, Aristóteles, Bateson o alguien más? Muchos enfoques en la investigación y la aplicación se orientan en función de otros, pero en realidad deberíamos proceder en función del objeto de nuestro interés epistemológico y menos en función de personas que pueden haber dirigido su atención a un objeto muy diferente. Y eso no tiene nada que ver con lo bien que se pueden clasificar sus declaraciones. Pero no importa cómo, quién o hacia qué dirijamos nuestro interés científico, siempre vuelven a nosotros algunas figuras, o como el psicólogo suizo del desarrollo y pensamiento Hans Aebli (1980, p.83) comentó acertadamente (en un contexto completamente diferente, por cierto), "Es cierto que uno nunca se mete dos veces en el mismo río. ¡Pero las escenas de baño son las mismas!"

1.1 El mundo es relativo

La filosofía de la ciencia puede entenderse según el siguiente lema:

El mundo es relativo y eso a su vez no cambia.

Todas las observaciones siguientes sobre la filosofía de la ciencia, así como sobre los métodos de investigación cuantitativos, cualitativos y lógicos, se guían por este lema. Resume todas las opiniones y puntos de vista presentados y debatidos en este libro. Sin embargo, la mutabilidad del mundo y la cierta falibilidad de los modelos no significan arbitrariedad, sino que exigen apertura y una actitud crítica y humorística hacia

nuestros propios puntos de vista, que normalmente sabemos defender con vehemencia. El resultado es que la ciencia no es necesariamente siempre el medio de elección. Por otro lado, requiere que la ciencia desarrolle instrumentos metodológicamente controlados y reproducibles que puedan aplicarse según criterios discutibles. Para ello, sin embargo, necesitamos *una teoría sobre la ciencia*: "por qué, por qué", por un lado, y "cuándo, cómo, cuándo no, cómo no", por otro.

> Recordatorio 1.1: Trabajo científico
>
> El trabajo científico se basa siempre en modelos y el mundo es cambiante. Los modelos son per se erróneos y, por tanto, nunca se corresponden exactamente con la realidad o la "verdad". Pero muchos modelos son extremadamente útiles para describir el mundo, de forma cuantitativa, cualitativa, lógica o combinada. A veces incluso son capaces de describir la variabilidad del mundo tan bien que las explicaciones, las predicciones y las intervenciones sean posibles de forma que puedan relacionarse directamente con la realidad.

1.2 Conocimiento, comprensión y sabiduría

Según Groeben y Westmeyer (1981), las tareas de la ciencia son

- *Explicación* - de acontecimientos pasados
- *Pronóstico* - predicción basada en toda la información disponible hasta la fecha.
- *Tecnología* - intervenir en el aquí y ahora para actuar a partir de la información disponible y el conocimiento de los posibles rumbos futuros, de modo que se posibilite la evolución más favorable.

La filosofía de la ciencia aborda la cuestión de las reglas y criterios primordiales según los cuales la propia ciencia genera conocimiento y está estrechamente vinculada a la cuestión de la capacidad cognitiva humana. Sin embargo, también se ocupa de las interacciones sociales dentro de la comunidad humana para crear algo nuevo juntos. Para esta investigación también son necesarios métodos científicos, a los que se aplican las mismas reglas que las que realmente se van a investigar a través de ellos. Así pues, la cuestión de un conocimiento "final" o "absoluto" nunca puede responderse (cuestión del huevo y la gallina), ya que uno trata de justificar al otro. Al hacerlo, el proceso prácticamente vuelve sobre sí mismo como fuente y surge un argumento circular. Los seres humanos carecemos de un punto de referencia absoluto para poder determinar dónde nos encontramos, comparable a una medida de diferencia de "éste es nuestro conocimiento" frente a "existe un conocimiento absoluto y ahí es donde queremos ir", y mucho menos para saber cómo avanzar y, por supuesto, en la dirección correcta.

En este sentido, la ciencia trabaja fundamentalmente con un criterio relativo de verdad. *Este dice que siempre podemos equivocarnos*, hagamos lo que hagamos. No podemos salir de este sistema aplicando rigurosamente métodos de trabajo científicos. Los criterios de verdad deben aplicarse a la propia ciencia. Esto significa que la filosofía de la ciencia como "ciencia sobre la ciencia" (metaposición) también sigue este criterio relativo del conocimiento. El mejor ejemplo, no siempre comprensible para el gran público, es el trabajo de Kurt Gödel (1906-1978) sobre las oraciones formalmente incompletas. Ya Bertrand Russell (1872-1970) y Alfred North Whitehead (1861-1947) afirmaron en los Principia Mathematica sobre la teoría lógica de los tipos (Whitehead & Russell, 1910) que los elementos de un sistema no pueden juzgar a todo el sistema. Al fin y al cabo, son partes del sistema y, como tales, se sitúan en una posición relativa respecto a todos los demás elementos del sistema. Nunca son superiores a ellos. Estas ideas encontraron su camino

en los enfoques de terapia sistémica de la Escuela de Palo Alto (Watzlawick, Beavin & D. D. Jackson, 1974), entre otros, y al mismo tiempo sirvieron como base de la teoría de la comunicación de Gregory Bateson (1985). Y nada más - aunque con otras palabras - sugiere la Alegoría de la caverna de Sócrates (Platón, Politeia, Libro VII, 514a, 1994): Para adquirir el "verdadero" conocimiento, no basta con interpretar e interpretar las sombras de la pared, sino que hay que levantarse, mirar hacia la "luz" y salir de la cueva. Esto implica experimentar y actuar, no sólo pensar e imaginar. Aplicado a la ciencia, esto significaría que, para obtener un conocimiento absoluto de nuestro reino, tendríamos que abandonar por completo el reino de la ciencia y mucho más - el de la mente y la materia - y con él el universo sensual que conocemos, por decirlo en pocas palabras.

Eso no es tan fácil. E incluso si lo consiguiéramos, careceríamos por definición de instrumentos como la lengua. Incluso si lo consiguiéramos, careceríamos por definición de los instrumentos, como el lenguaje, para comunicar adecuadamente los conocimientos adquiridos de este modo.

1.3 Los conocimientos dependen de la cultura

De la comparación cultural aprendemos también que el criterio científico de la verdad que es válido aquí en nuestro país - en general, la creencia en el pensamiento lógico, las razones y los instrumentos precisos (físicamente manifiestos), así como los análisis, depende de la cultura y no representa la forma más elevada de conocimiento en todo el mundo. Por ejemplo, si se utilizan las enseñanzas prácticas del último Buda, Siddhata Gotama (500 a.C.; Schumann, 1999), la cognición se descompone en las tres categorías de *sutta-maya paññā*, *cinta-maya paññā* y bhāvanā-mayā paññā (Goenka, 1991):

- *sutta-maya paññā* - oído y creído.
- *cinta-maya paññā* - intelectualmente autoconcluido
- *bhāvanā-maya paññā* - experimentado directamente en uno mismo

Sin embargo, el conocimiento en su forma completa no surge del intelecto, sino que abarca esencialmente y requiere necesariamente una conducta ética de la vida (*sīla*), el dominio de la propia mente (concentración, *samādhi*), y precisamente la sabiduría debida a la experiencia directa y la comprensión de la naturaleza de la mente y la materia (*bhāvanā-maya paññā*) dentro del propio fenómeno mente-cuerpo. El resultado es una actitud benevolente hacia todos los seres. En este proceso, las sensaciones corporales normales (vedanā) desempeñan un papel especial en la obtención de la percepción, como se afirma en el Mahā satipat.thāna nasuttam. (V.R.I, 1993), uno de los discursos doctrinales más importantes del budismo Theravāda.

Desde este punto de vista, la ciencia tal como la entendemos no es la forma más elevada de conocimiento. El hecho de que el intelecto no transmita la forma más elevada posible de cognición es ciertamente difícil de comprender en un principio para muchos científicos. Sin embargo, no debemos descartar esta idea sin más. Al fin y al cabo, se basa en la práctica y no sólo en la teoría, y tiene miles de años de antigüedad.

La sabiduría de Buda se produce cuando la mente simplemente reconoce y abandona la ilusión de una existencia propia e inmutable, y ve las cosas como realmente son: cambiantes, sufrientes e insustanciales, consistentes en relaciones de causa-efecto. Se trata de un camino espiritual y a la mayoría de los científicos les resultaría muy difícil distinguir la ciencia de él. Otra limitación es que dicho conocimiento es, por definición, de naturaleza subjetiva y depende del grado de perspicacia y desarrollo individual, que no puede acelerarse a voluntad. Aquí es donde la mayoría de los científicos, y especialmente los científicos naturales, probablemente abandonen, ya que la subjetividad se considera casi "maligna" para los enfoques de la ciencia, en su mayoría orientados a los números. Por otra parte, muchos científicos (naturales) de la historia no dejan de especular sobre Dios o incluso de referirse a él. Entre ellos se encuentran Charles Darwin, Albert Einstein, Werner Heisenberg, Thomas A. Edison o Max Planck, es decir, la élite de la ciencia (natural). Quienes especulan sobre Dios no deberían tener miedo de explorar sus propias mentes.

Sin embargo, desde la perspectiva del camino espiritual, nuestros descubrimientos científicos serían prácticos y útiles, pero carecerían de importancia si se consideran en términos absolutos, porque en realidad no se trata de eso. Todo es cuestión de perspectiva: la fe desempeña un papel en este sistema en la medida en que la fe es importante cuando se basa en la realización interior directa de uno mismo. El Buda veía su propio sistema lejos de las creencias religiosas, sino como una investigación científica intrasubjetiva verificable de la mente y la materia, y eso debería al menos darnos que pensar. A raíz de ello, se fundaron universidades budistas, a menudo centradas en el pensamiento lógico científico (por ejemplo, la Universidad de N*a*land*a*, cerca de Rajgir, en el estado indio de Bihar, 500-1200, el mayor centro de enseñanza del mundo antiguo). Incluso hoy en día, la lógica y las habilidades de debate se consideran pilares centrales de la educación en los monasterios budistas. Esto se ejemplifica con las posibilidades mínimas de decisiones lógicas, que pueden ir mucho más allá del pensamiento dualista, como el tetralema.

Con el tetralema, existe "el uno", "el otro", "ambos", "ninguno" o "algo completamente diferente". El alumno de Stegmüller, Varga von Kibéd (Sparrer & von Kibéd, 2000), utilizó esta lógica como una forma de trabajo de constelación estructural ("tetralema") para su trabajo de constelación sistémica. La dialéctica hegeliana, con el papel central de la negación y posteriormente de la síntesis, también persigue el objetivo de llevar el conocimiento a un nivel superior y se sitúa así en línea directa con Sócrates y Platón.

La fe siempre desempeña un papel implícito en la ciencia. Hoy en día, por ejemplo, existe una miríada de hallazgos científicos que, a nivel de artículos de revistas, suelen aparecer como *la unidad publicable más pequeña*. Aquí es difícil no perder la visión de conjunto. Y no suele haber tiempo suficiente para las réplicas propias. Así que uno se cree las conclusiones de los demás. Pero no todos los estudios empíricos se han realizado y se realizan siempre correctamente. Casi ningún estudio se replica y a veces no sólo cabe sospechar que los resultados científicos han sido manipulados. Reconstruir todo esto, aunque sea minuciosamente, no es factible en la vida científica cotidiana. Así que los resultados de las publicaciones se creen de forma "críticamente apreciativa". Sin embargo, no se experimentan en absoluto, y a menudo el tema ni siquiera se comprende del todo intelectualmente, sino que sólo se utilizan partes. Fuera del propio campo de especialización, en cualquier caso es difícil evaluar correctamente la gravedad de los resultados o incluso separar lo que es relevante. Una excepción es el procedimiento metodológico (incluidos los análisis de datos), que a menudo se formula de forma interdisciplinar y permite así una estimación aproximada de los resultados de los estudios. Una sólida formación metodológica es, por tanto, uno de los pilares de la ciencia.

Ya podría darse un paso intermedio y una apertura incluyendo explícitamente la intuición como factor importante de la ciencia. La intuición como percepción directa e inmediata ha desempeñado hasta ahora un papel más en la filosofía y fue entendida por Carl Gustav Jung (1875-1961) como una función psicológica básica que media las percepciones del individuo y llega hasta el inconsciente (colectivo). Desde un punto de vista cognitivo, la intuición es una capacidad de tratamiento especial de la información, en particular de complejidad. También en este caso se utiliza un puente de la conciencia a la inconsciencia por medio de la intuición. Hoy en día, la intuición desempeña implícitamente una función importante en las teorías de gestión, pero también en el diseño de software y otros ámbitos. Así, se observa que las sustancias desempeñan un papel en el fomento de la creatividad y la intuición en el trabajo, además del rendimiento, la productividad y la concentración.

Un ejemplo es el creciente consumo de dosis de micro-LSD en el desarrollo de software en Silicon Valley. Así pues, parece legítimo conceder explícitamente a la intuición un lugar en la ciencia y fomentar su desarrollo y modelizarla adecuadamente en la teoría científica. La abducción (véase más adelante) según Charles Sanders Peirce (1839-1914) incluye la intuición como procedimiento de inferencia epistemológica. Esto queda claro cuando Peirce (1965, CP 5.181, p.113) comenta el proceso real del razonamiento abductivo:

> „The abductive suggestion comes to us like a flash. It is an act of insight, also of extremely fallible insight. It is true that the different elements of the hypothesis were in our minds before; but it is the idea of putting together what we had never before dreamed of putting together which ashes the new suggestion before our contemplation.“ ["La sugerencia abductiva nos llega como un flash. Es un acto de perspicacia, también de extrema perspicacia falible. Es cierto que los distintos elementos de la hipótesis estaban antes en nuestra mente; pero es la idea de juntar lo

que nunca antes habíamos soñado juntar que ceniza la nueva sugerencia ante nuestra contemplación".]

¿Qué es este destello sino una intuición? Una intuición que hay que poner en palabras, modelar científicamente y verificar. Sin embargo, la adición de la intuición o del conocimiento superior no significa que debamos apartarnos de la ciencia y de sus elementos indispensables. Por el contrario, la ciencia nos ofrece la máxima precisión accesible y comunicable cuando se trata de pensar (lógicamente), analizar datos o reconstruir y deducir causas y conclusiones para la práctica. De lo contrario, una llamada de móvil no funcionaría, no habría televisión y no podríamos ayudar a los drogadictos o a las personas traumatizadas, ni mostrar cuál es la mejor manera de aprender en la escuela y en qué condiciones. Del mismo modo, no podríamos averiguar sin la ayuda de las estadísticas que el clima de la Tierra se está calentando o en qué condiciones el acoso sexual es especialmente frecuente o que la creciente brecha entre ricos y pobres no sólo existe, sino que va en aumento. No sólo nuestro bienestar, lujo o conocimiento placentero dependen de la ciencia, sino que la ciencia contribuye significativamente a nuestra vida y supervivencia en la Tierra, pero también a lo contrario (por ejemplo, el desarrollo de bombas atómicas, el cambio climático, la manipulación de las personas, etc.).

Cuando se trata de comprensión y conocimiento, especialmente en las ciencias naturales y las matemáticas, el físico Jaynes (2003, p.20) señala

„Perhaps we have here the reason why science and mathematics are the most successful of human activities: they deal with propositions which produce the simplest of all mental states. Such states would be the ones least perturbed by a given amount of imperfection in the human mind.“

En las ciencias sociales, nos enfrentamos a estados mentales mucho más complejos. La modelización de los seres humanos y, en especial, de las interacciones sociales de individuos o grupos es, por consiguiente, difícil, aunque puedan modelizarse matemáticamente. Empecemos, pues, por las herramientas del oficio: el pensamiento estructurado y el razonamiento.

Capítulo 2
Razonamiento Lógico en el Proceso de Investigación

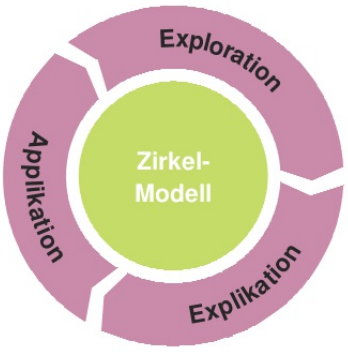
> »Contrariwise, continued Tweedledee, if it was so, it might be; and if it were so, it would be: but as it isn't, it ain't. That's logic.«
>
> *Through the Looking Glass*.
> Lewis Carroll, 1871

El punto de partida de los estudios en las ciencias sociales suele ser el hecho de que los investigadores se encuentran con fenómenos hasta entonces inexplicados, inesperados o problemáticos en los campos de acción social. A continuación, intentan encontrar explicaciones a las mismas, hacer posibles predicciones para situaciones similares o elaborar recomendaciones para resolver problemas. Estos esfuerzos definen el espacio de un proceso, que puede dividirse en tres sectores (véase la Fig. 2).

- Aplicación: el campo de acción social que se va a estudiar,
- Exploración: el sector de la elaboración de preguntas de investigación y el diseño de la investigación
- Explicación: el sector de la búsqueda de explicaciones y fundamentos del fenómeno crítico.

Especialmente cuando se trata de esclarecer problemas, también se buscan planteamientos de soluciones, que luego pueden aplicarse a su vez en el campo de acción social, el sector de aplicación. Se puede profundizar a voluntad (véase la Fig. 2.2). Todo esto tiene lugar en el contexto de los objetivos de la ciencia enumerados anteriormente (Groeben & Westmeyer, 1981). Si consideramos que los problemas ocurren en el presente, pero sólo tenemos conocimientos del pasado y buscamos soluciones para el futuro, siempre se requieren conclusiones plausibles:

- Puede obtener información sobre actos anteriores. Estos acontecimientos y su ocurrencia en el pasado deben explicarse para poder comprenderlas.
- A partir de ahí, queremos derivar predicciones para el futuro y la evolución futura teniendo en cuenta diversas influencias.
- Y, por supuesto, queremos diseñar posibles acciones en el presente para que tengan un efecto favorable en el presente.



Aplicación
Exploración
Explicación

Fig. 2.1. *Modelo circulo del proceso de investigación*

En resumen, se trata de resolver o reducir los problemas de los que partió la investigación y de hacer más probable una evolución deseable.

Vemos que en cada una de estas tres áreas, el razonamiento plausible desempeña un papel central a la hora de cumplir los requisitos de generar hipótesis y teorías, desarrollar un diseño de investigación y diseñar intervenciones para la práctica. Para ello, el primer paso es formular nuestros supuestos, presunciones y consideraciones teóricas de forma lógica, coherente y sin contradicciones. De este modo, la posterior recogida y análisis de datos no debería fallar porque los cimientos de nuestra investigación fueran débiles. La ciencia necesita un buen trabajo teórico antes de recoger cualquier dato. O dicho de otro modo: necesitamos un razonamiento estable que muestre el camino desde la información disponible hasta la evaluación de la información - es decir, su interpretación - de forma comprensible para los demás. Las interpretaciones no son arbitrarias ni aleatorias, sino que los datos se controlan metódicamente, es decir, se interpretan con sentido y siguiendo unas reglas. El control metódico nos permite comunicar, transmitir y enseñar, así como comprobar el progreso del aprendizaje y asesorar en caso de ambigüedades.

En un segundo paso, las interpretaciones y las decisiones resultantes también deben contextualizarse de forma significativa y basada en normas. La orientación de la norma es especialmente importante en este caso, porque ni las interpretaciones ni las decisiones pueden depender de una persona. Las conclusiones de distintas personas que hagan uso de las mismas reglas, construcciones de significado e información deben ser lo más congruentes posible, si no idénticas. Esto no significa que estas personas tengan que llevar a cabo cada paso del análisis de forma idéntica: es posible y está permitida una gran variabilidad de enfoques. Las conclusiones finales deberían apuntar a lo mismo.

A modo de ejemplo, imaginemos un grupo de pedagogos que han observado en el ámbito de la escuela que algunos alumnos no se esfuerzan, aunque sin duda tendrían la oportunidad de mejorar su rendimiento escolar mediante un aprendizaje más intensivo. La pregunta de la investigación es, por tanto, qué explicaciones puede haber para este fenómeno.

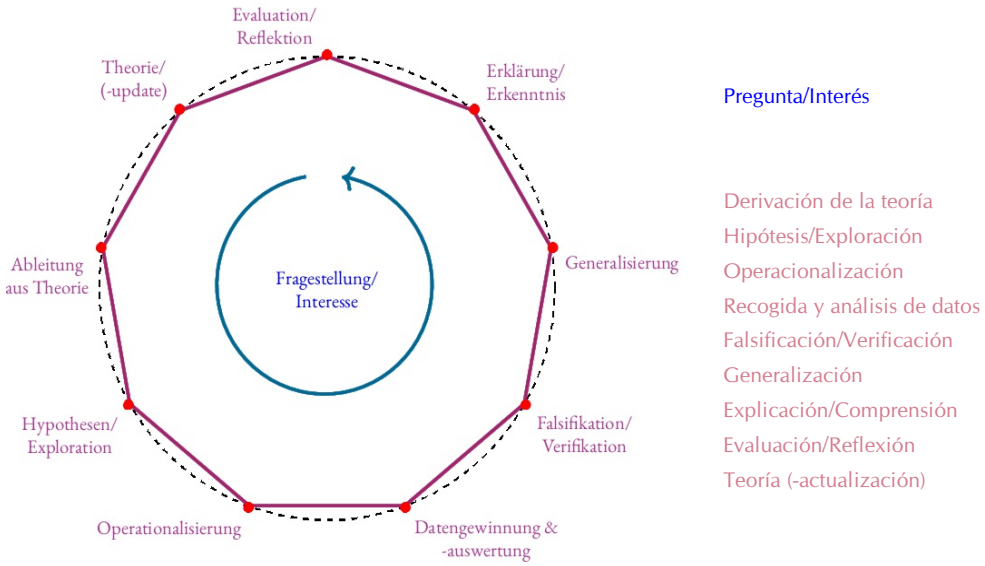


Fig. 2.2. *Profundización del proceso de investigación cíclica*

En otras palabras: el grupo aún no tiene hipótesis que pueda poner a prueba, pero el primer paso es generar hipótesis plausibles.

Al hacerlo, los investigadores intentan, como es natural, extraer conclusiones sobre las posibles causas del fenómeno a partir de las circunstancias del campo de acción social (falta de esfuerzo, posibilidades de mejora, etc.) y de los hallazgos científicos disponibles. En general, se distinguen tres formas abstractas de inferencia: inducción, deducción y abducción.

La tabla 2.1 muestra su relación. Para entenderlas, veamos primero los tres componentes básicos que suelen aparecer en las conclusiones lógicas e ilustrémoslos con un ejemplo bien conocido en lógica (véase el cuadro 2.1).

1. Premisa principal, ley o regla (generación): "Cuando llueve, la carretera está mojada".
2. Premisa secundaria o caso concreto: "La calle de enfrente de mi casa está mojada".
3. Conclusión (inferencia) o resultado: "Está lloviendo (o ha llovido)".

Peirce (1965, CP 5.171) describe estas tres formas de inferencia de la siguiente manera:

- „Deduction proves that something must be;
- induction shows that something actually is operative;
- abduction merely suggests that something may be.“

Tabla 2.1: *Formas de razonamiento lógico*

	Inducción
Caso	Se da B
Conclusión (resultado)	B tiene características según la regla A
Regla (generación)	La regla A se aplica a (todas) B
	Deducción
Regla (existente)	Si A, entonces B
Caso	ahora A
Conclusión (resultado)	por tanto B
	Abducción
Conclusión (resultado)	Observación A (o B)
Regla (generación)	(todas) B están después de A (o A después de B)
Caso* (aplicación de la regla)	A y B están relacionados (si A, entonces B resp. v.v.)

* debe probarse empíricamente

2.1 La conclusión inductiva

El primer modo señalado en la tabla 2.1, el modo de *inducción*, es muy común en los estudios científicos. Se dispone de un caso dado ("La calle está mojada") y una regla general ("Si llueve o ha llovido, entonces la calle está mojada"). Ahora se saca la conclusión "Está lloviendo", es decir se concluye hacia la existencia de un componente "Si ..." general, formulado más abstractamente: *utilizando la inducción se concluye desde*

algo particular hacia algo general. El problema es que la conclusión inductiva basado en la regla general podría ser falsa – y considerando todas las circunstancias muchas veces es falsa. Podría ser que una tubería de agua reventó o los camiones esparcidores de la limpieza callejera pasaron por la casa o, simplemente, las gafas del observador están mojadas y la calle solo parece mojada. Por consiguiente, hay cada vez una cierta posibilidad o probabilidad de que la conclusión inductiva fracasa. En el departamento "Métodos de inferencia" de la estadística se trata precisamente de calcular la probabilidad de tales conclusiones inductivas basadas en datos limitados.

Se discute en la ciencia el problema de la inducción desde hace muchos siglos (vease p.ej. Stegmüller, 1975); particularmente el filósofo escocés David Hume (1711–1776) se dedicaba a las preguntas vinculadas con la inducción. En nuestra vida cotidiana utilizamos casi siempre la inducción como modo de generalización. ¡Sacamos permanentemente conclusiones inductivas! Por ejemplo partimos de la base de que un hombre (salvo niños pequeños) que ha mentido una vez va a mentir otra vez. De la misma manera se suponía hasta hace muy poco tiempo que mujeres solteras con niños llevan una vida sospechosa. Basado en observaciones particulares que hacían los campesinos sobre el tiempo y las temperaturas estos han concluido inductivamente pronósticos climatológicos (que se han resumido en un "almanaque agrícola"). Otro ejemplo: ya que Donald Trump (el antiguo presidente de los EE.UU.) suele difundir sus opiniones por Twitter con un vocabulario fácil y una cantidad limitada de palabras concluimos que él no puede comunicar de otra manera. ¿Es así? ¿No podría ser simplemente una estrategia para dirigirse a determinados grupos de personas de una determinada manera? O incluso ¿ambas posibilidades? Como ponen de manifiesto estos pocos ejemplos, nuestra inducción cotidiana es enormemente defectuosa. Nunca pedimos lo contrario y no definimos nuestras suposiciones con precisión fundada, sino sobre todo estereotipada. ¿Hay mujeres casadas que también viven inmoralmente? ¿Y cómo o a partir de qué se define realmente la moralidad? ¿Por qué hay diferencias fundamentales entre hombres y mujeres y por qué se aplican normas diferentes? La inducción es una parte muy común de nuestra vida cotidiana, pero ¿podemos por tanto confiar en ella para llegar a conclusiones generalmente válidas? (Gürtler y Huber, 2005). A continuación nos gustaría demostrar que en las discusiones científicas se discute a menudo exactamente sobre este punto, pero la pregunta en sí es en realidad errónea.

Antes de profundizar en el problema del error en el razonamiento inductivo (Carnap & Stegmüller, 1959; Stegmüller, 1973a, 1973b, 1975), veremos ejemplos de Pólya (1954a, p.11) en los que un lógico, un matemático, un físico y un ingeniero comentan el razonamiento de los demás.

> "Mira a este matemático", dijo el lógico. "Observa que los primeros noventa y nueve números son menores que cien e infiere de ahí, por lo que llama inducción, que todos los números son menores que cien".
>
> "Un físico cree", dijo el matemático, "que 60 es divisible por todos los números. Observa que 60 es divisible por 1, 2, 3, 4, 5 y 6. Examina algunos casos más, como 10, 20 y 30, tomados al azar, como él dice. Como 60 también es divisible por éstos, considera que la prueba experimental es suficiente."
>
> "Sí, pero mira a los ingenieros", dijo el físico. "Un ingeniero sospechaba que todos los números impares son números primos. En cualquier caso, el 1 puede considerarse un número primo, argumentó. Luego vienen el 3, el 5 y el 7, todos indudablemente primos. Luego viene el 9; un caso incómodo, no parece ser un número primo. Sin embargo, el 11 y el 13 son ciertamente primos. "Volviendo al 9", dijo, "concluyo que el 9 debe ser un error experimental".

Formalmente (véase la Tab. 2.1), la figura de la conclusión inductiva puede representarse tradicionalmente como sigue y - utilizando el ejemplo "la carretera está mojada" - pueden postularse otras reglas potencialmente válidas además de la regla inductiva, que se basan en premisas secundarias diferentes (véase la Tab. 2.2). Como indican los signos de interrogación en la conclusión o en la premisa secundaria, no es posible identificar ninguna regla sin duda válida. Esto, a su vez, puede asumirse como una regla válida, al menos para las ciencias sociales, porque sencillamente no hay leyes deterministas en este ámbito: Las personas y las interacciones humanas son demasiado complejas y claramente más complicadas que las "simples" leyes físicas de la naturaleza.

E incluso en ese caso, habría que preguntarse cómo pueden formularse leyes deterministas que igualen, por ejemplo, los estrictos requisitos de la inducción matemática. No hay que olvidar que en el nivel de la

mecánica cuántica sólo existen probabilidades y que faltan por completo las supuestas certezas que tal vez siguen presentes como una ilusión en nuestro nivel cotidiano. Y a continuación incluso estas probabilidades dejan de existir bajo la límite de la longitud de Planck, ya que por definición esta zona ya no se puede explorar.

Tabla 2.2: *Inferencia inductiva*

	Abstracción formal	**Ejemplo**
Caso	El caso dado es B.	La carretera está mojada.
Conclusión (Resultado)	[???] B tiene características según la regla A.	Está lloviendo.
Premisa secundaria 1	[???] Sin embargo, una regla C, D o E podrían ser válida.	Los cristales están mojados, los barrenderos no estaban, se ha roto una tubería, etc.
Premisa secundaria 2	[???] Sin embargo, una regla C, D o E podrían ser válida.	La calle está seca.
Regla (Generación)	La regla A se aplica a (todas) las B.	Si llueve, la carretera está mojada.
Reglas alternativas	La regla C se aplica a (todas) B.	Cuando se mira a través de gafas mojadas, todo parece mojado.

A lo largo de la historia, varios autores han intentado formular la regla inductiva en términos probabilísticos. En este sentido, cabe destacar la obra del filósofo y representante de la lógica empirismo, Rudolf Carnap (1891-1970). Sin embargo, Carnap fue atacado y criticado desde varios flancos por estos intentos (incluso por Karl Popper), lo que dio lugar a una amplia controversia (Michalos, 1971). A la inversa, Carnap consideraba que el deductivismo puro de Popper era inviable para las teorías y los problemas científicos genuinos, ya que no daría cuenta adecuadamente de la naturaleza probabilística de las cosas. En su lugar, defendió la idea de que las teorías bien confirmadas empíricamente tienen una mayor probabilidad (Carnap & Stegmüller, 1959; Carnap, 1973). De este modo se cuantifica el grado de confirmación. El punto de inflexión es siempre el intento de introducir un grado de confirmación calculable para las teorías científicas. Muchas de estas ideas pueden encontrarse en el actual estadística de Bayes. El matemático húngaro George Pólya (1887-1985) también persiguió en sus obras el abandono de las leyes exclusivamente deterministas en el razonamiento inductivo (por ejemplo, 1954a, 1954b). Examinó muy de cerca las cifras e influencias del cierre potencial, especialmente en el caso del cierre por inducción. Su enfoque no matemático, casi ético, del tema es correspon-dientemente sorprendente (Pólya, 1954a, p.7):

> „In our personal life we often cling to illusions. That is, we do not dare to examine certain beliefs which could be easily contradicted by experience, because we are afraid of upsetting our emotional balance. There may be circumstances in which it is not unwise to cling to illusions, but in science we need a very different attitude, the inductive attitude. This attitude aims at adapting our beliefs to our experience as efficiently as possible. It requires a certain preference for what is matter of fact. It requires a ready ascent from obervations to generalizations, and a ready descent from the highest generalizations to the most concrete observations. It requires saying 'maybe' and 'perhaps' in a thousand different shades. It requires many other things, especially the following three.
>
> First, we should be ready to revise any one of our beliefs.
>
> Second, we should change a belief when there is a compelling reason to change it.
>
> Third, we should not change a belief wantonly, without some good reason."

Según Pólya, la primera afirmación requiere "intellectual courage" (coraje intelectual), la segunda "intellectual honesty" (honestidad intelectual) y la tercera "wise restraint" (sabia restricción). A continuación, Pólya ofrece una solución basada en un *razonamiento plausible*.

La premisa principal inicialmente determinista "A implica B" señala que B es consecuencia de A , sin que exista necesariamente una relación lógico-causal. Tampoco hay pruebas de que esta relación sea siempre y en toda circunstancia válida. El convincente suposición de una relación lógico-causal sería, en consecuencia, una falacia lógica, a saber, que la regla A se aplica a B, siempre y en todas partes. La premisa secundaria representa la información disponible en función del contexto.

Si en un caso recibimos la información de que A es falso ("no llueve"), no sabemos necesariamente si B es VERDADERO o falso. El hecho de que "A implica B" sea cierto no significa necesariamente al mismo tiempo que "no-A implica no-B". Si recibimos la información "B es VERDADERO" ("la carretera está mojada"), esto tampoco indica necesariamente que A ("está lloviendo") deba ser VERDADERO. De hecho, no excluye la posibilidad de que también se cumpla otra regla "C implica B", etc. En tales casos, sólo disponemos de cierta evidencia inductiva para nuestras conclusiones ("A es falso/ la carretera está seca" o "B es VERDADERO/ la carretera está mojada"). Sólo podemos suponer cierta credibilidad a las relaciones respectivas, que se ve incrementada por el éxito de las pruebas empíricas. Pólya (1954a) habla aquí de una expresión de racionalidad cuando suponemos, a partir de la información "B es VERDADERO", que aumenta la plausibilidad o credibilidad de "A es VERDADERO". Lo mismo se aplica a una mayor probabilidad de "B es falso" si "A es falso" es cierto. Sin embargo, inicialmente no sabemos nada sobre el alcance exacto del cambio en el grado de probabilidad con respecto a la validez y credibilidad de la regla. Las tablas 2.4 y 2.3 ejemplifican respectivamente el procedimiento inductivo de Pólya con el objetivo de cambiar la credibilidad y la plausibilidad respectivamente. Como muestran los distintos patrones de razonamiento enumerados, el espectro entre lo VERDADERO y lo falso se cubre con sus distintos matices de grises y sombras. Esto se aplica tanto a las premisas como a las inferencias plausibles. El patrón que recorre todos estos ejemplos abstractos es el de la anulación y el refuerzo mutuos a través de la complementariedad. Si ambas premisas son inequívocas (VERDADERO, falso), el ejemplo adquiere el estatus de "demostrativo" en el sentido de Prueba y surge una clara conclusión plausible (nº 1). Si la relación entre A y B es clara ("A implica B"), pero B en sí es menos creíble, esto se contagia a A y A también pierde credibilidad (nº 3). Si cambia la relación de A y B (columna "premisa 1"), cambia la dirección de los argumentos y las conclusiones y, por tanto, los grados de plausibilidad. Así, pueden examinarse consecuencias, repeticiones de éstas, diferentes grados de probabilidad, analogías, causas o exclusiones mutuas (véanse los subapartados de las Tab. 2.3 o 2.4, traducidos según Pólya, 1954b, p. 4 ss. y p. 26, Tab. 1). Por ejemplo, en los números 8 y 9, Pólya (1954b, p. 9, cursiva en el original) señala

> „The verification of a consequence counts more or less according as the consequence is more or less improbable in itself. The verification of the most surprising consequences is the most convincing."

Existe, pues, un sistema básico para relacionar unas premisas con otras de forma más o menos compleja. manera y obtener como resultado un argumento racional plausible. En la literatura, los términos credibilidad, fiabilidad o persuasión se utilizan como sinónimos de verosimilitud (Studer, 1996). o persuasión (Studer, 1996b, p.7). Jaynes (2003, p.5) formaliza este proceso y postula un "degree of pausibility" (grado de plausibilidad), lo contextualiza y describe la asignación de números para evaluar dicha plausibilidad. Studer (1996b, p.9) traduce esta evaluación de la plausibilidad como "grado de creencia razonable" siguiendo a Jeffreys (1939/1961, "degree of reasonable belief"). Lo importante es que estos tipos de racionalidad tienen lugar fundamentalmente en un contexto y nunca existen al margen de él, sino siempre según reglas concretas que pueden modelarse. Por tanto, no se aplica la objeción de Popper, que se refiere a la comparación de normas descontextualizadas. Aquí, sin embargo, nos ocupamos de la comparación de normas contextualizadas. En relación con las asignaciones numéricas, esto significa (véase el capítulo 6.2) que existen probabilidades condicionales y no incondicionales. Indican la probabilidad de que en un contexto se considere VERDADERO (o falso) el contenido de una afirmación (proposición). En la estadística bayesiana, el álgebra de Boole está vinculada a la teoría de la probabilidad (véase el capítulo 6). Si recapitulamos los supuestos básicos de la inducción, nos encontramos con Jaynes (2003). En su libro publicado póstumamente

sobre "Probability Theory as Logic", editado por su alumno G. L. Bretthorst, Jaynes (ibid., p. 326) señala los falsos supuestos básicos que subyacen a la cuestión del problema de la inducción:

Tabla 2.3 & 2.4: *Conclusión inductiva según Pólya (1954b)*

No.	**Premisa 1**	**Premisa 2**	**Premisa 3**	**Conclusión plausible**	**Categoría**
Indagación de consecuencias					
1	A implica B	B falso		A falso	ostensiva
2	no-A implica no-B	no-B falso		no-A falso	ostensiva
3	A implica B	B menos creíble		A menos creíble	ostensiva sombreada
4	A implica B	B más bien creíble		A un poco más creíble	inductiva sombreada
5	A implica B	B de verdad		A de verdad	inductiva
Verificación eficaz de varios resultados					
6	A implica B_{n+1}	B_{n+1} parece mucho a las consecuencias B_1, B_2 ... B_n de A (ya verificadas)	B_{n+1} de verdad	A es un poco más creíble	inductiva sombreada ¿?
7	A implica B_{n+1}	B_{n+1} es muy distinta de las consecuencias B_1, B_2 ... B_n de A (ya verificadas)	B_{n+1} de verdad	A es mucho más creíble	inductiva ¿?
Verificación de un resultado improbable					
8	A implica B	B es más bien probable per se	B de verdad	A es un poco más creíble	inductiva sombreada ¿?
9	A implica B	B es más improbable per se	B de verdad	A es mucho más creíble	casi completa-mente inductiva ¿?
Conclusión por analogía					
10	A análogo a B	B más bien creíble		A un poco más creíble	inductiva sombreada ¿?
11	A análogo a B	B de verdad		A más bien creíble	inductiva ¿?
Examinar una causa posible					
12	A implica B	B de verdad		A de verdad	ostensiva
13	A implica B	B más bien creíble		A más bien creíble	ostensiva sombreada
14	A implica B	B menos creíble		A un poco menos creíble	inductiva sombreada
15	A implica B	B falso		A falso	inductiva
Examinar una suposición contradictoria					
16	A incompatible con B	B de verdad		A falso	ostensiva
17	A incompatible con B	B más bien creíble		A menos creíble	ostensiva sobreada

18	A incompatible con B	B menos creíble		A un poco más creíble	inductiva sombreada
19	A incompatible con B	B falso		A más bien creíble	inductiva

„As we noted ..., some philosophers have rejected induction on the grounds that there is no way to prove that it is 'right' (theories can never attain a high probability); but this misses the point. The function of induction is to tell us not which predictions are right, but which predictions are indicated by our present knowledge. If the predictions succeed, then we are pleased and become more confident of our present knowledge; but we have not learned much."

Epistemólogos y filósofos como David Hume o Karl Popper (1902-1994) negaron por completo la posibilidad de generalización a partir de inferencias inductivas. Jaynes (2003, p. 310 ss.), por su parte, sostiene que todo el conocimiento científico en sentido amplio puede remontarse a la inducción. Además, se pone de manifiesto un error lógico de pensamiento, sobre todo en el caso de Popper. Según Jaynes, no se trata de utilizar la inducción para concluir la superioridad de una teoría en comparación con un número infinito de todas las teorías posibles y demostrar definitivamente esta superioridad. Se trata más bien de la verosimilitud de una teoría considerada en relación con un conjunto finito de explicaciones alternativas competidoras especificadas. Así, una hipótesis puede alcanzar una probabilidad muy alta o muy baja dentro de una clase de alternativas bien definidas. Por otra parte, la probabilidad en el espacio de todos los posibles concebibles, es decir, un conjunto infinito de teorías no está definida en absoluto y, por tanto, no tiene mayor relevancia para la práctica concreta de la investigación.

Aquí Jaynes se refiere a la inferencia en el sistema metodológico de la estadística bayesiana y concluye que la inferencia bayesiana se refiere a problemas determinados claramente definidos, y no, como Popper, a problemas indefinidos. Lo que Jaynes señala, pues, es la contextualización de una teoría o hipótesis ya mencionada. Las teorías siempre se ponen a prueba en un contexto definido de forma clara e inequívoca y si una de ellas resulta ser superior en términos de alta verosimilitud (probabilidad), ello es consecuencia de la comparación directa con las teorías competidoras relacionadas con el mismo contexto. Es incluso mejor no contrastar las teorías entre sí, sino organizar todos los casos especiales inclusivos en una teoría global. El objetivo de lograr la superioridad de una teoría contextualizada sobre un conjunto infinito de teorías indefinidas pero concebibles carece de sentido y no es muy científico. Esta es precisamente la razón por la que Jaynes critica el hecho de que las discusiones filosóficas sobre el problema de la inducción se realicen normalmente de forma lógica abstracta y no se relacionen con ejemplos y contextos reales. En cualquier estudio de un caso real, enseguida queda claro que no hay ningún contexto en el que exista un conjunto infinito de hipótesis razonables que compitan entre sí. Siendo realistas, es probable que sólo existan unos pocos y que éstos tengan probabilidades condicionales.

El *grado de plausibilidad* se revela así como una probabilidad condicional p de una proposición. La inducción se formula así como una cuestión de *plausibilidad condicional* y no como una afirmación de validez general. De este modo, las conclusiones no se generalizan de forma generalizada y descontextualizada, sino que se refieren siempre a situaciones concretamente existentes. Históricamente, el problema de la inducción gira precisamente en torno a esta cuestión: ¿es posible una generalización *siempre* válida de los hallazgos? Desde el punto de vista de los defensores de la plausibilidad (Pólya, Jaynes, Cox, etc.), esta pregunta no se responde con un sí o un no, sino que se afirma que esta pregunta no tiene sentido porque los hallazgos siempre están contextualizados. La plausibilidad es, por tanto, la del intento de explicación E sobre otros intentos D, F o G.

Los problemas reales no conocen un espacio infinito de posibles explicaciones teóricas. En la mayoría de los casos, sólo hay un pequeño número de explicaciones posibles cuyo poder explicativo puede determinarse empíricamente en comparación directa entre sí en forma de probabilidad condicional. El estadístico estadounidense Andrew Gelman, por su parte, va un paso más allá y sugiere que, en lugar de contrastar las teorías entre sí (por ejemplo, mediante factores de Bayes, véase el capítulo 6.8.1), se intente más bien integrar estas teorías y sus perspectivas en un modelo más complejo y determinar contextos para que, por ejemplo, los sucesos raros, los fenómenos marginales, los valores atípicos, etc. queden cubiertos

por una explicación teórica más compleja que incluya condiciones límite. Gelman favorece así un modelo fundamentalmente dialéctico, pero sin expresarlo de este modo.

En un ejemplo ficticio, nos preguntaríamos, por ejemplo, si podemos extraer conclusiones sobre todos los demás niños y su comportamiento en todas las situaciones de rendimiento imaginables a partir de la observación de la evitación del esfuerzo en un solo niño o en un grupo de niños. Inmediatamente queda claro que en una situación de problema científico real nadie se planteará seriamente una pregunta así. Más bien, nos preguntamos por las condiciones en las que pueden surtir efecto explicaciones como la protección de la autoestima, la atribución, etc. En lugar de ponerlos a prueba unos contra otros a lo "¿Qué es entonces?", podemos intentar integrar estos enfoques entre sí de tal forma que surja un modelo predictivo razonable y sólido, que por cierto pueda proporcionar intervenciones educativas adaptativas y no excluya los casos límite.

Aunque es imposible contrastar todas las teorías imaginables debido a la cantidad potencialmente infinita de teorías, sí se pueden probar, modelizar o examinar los cambios bajo la condición de una referencia contextual concreta. Así, una teoría puede resultar más explicativa que otra. Partiendo de la pregunta puramente retórica de Popper de si es posible una conclusión racionalmente justificada basándose únicamente en la observación de sucesos repetidos, Jaynes afirma que sólo puede responderse claramente con un no bajo la condición de mala información. Sin embargo, si en el sentido del razonamiento plausible se añade información previa, es decir, información contextual actual o información disponible del pasado, y este estado de conocimiento apunta a una conexión lógica entre los diversos acontecimientos observados, entonces, según Jaynes, la inducción es muy posible. Curiosamente, esta opinión cuenta con el apoyo del estadístico frecuentista Ronald A. Fisher, que escribe en su obra principal "El diseño de experimentos" (Fisher, 1935/1973, p.7):

> El grado de plausibilidad se revela así como una probabilidad condicional p de una proposición. La inducción se formula así como una cuestión de plausibilidad condicional y no como una afirmación de validez general. De este modo, las conclusiones no se generalizan de forma generalizada y descontextualizada, sino que se refieren siempre a situaciones concretamente existentes. Históricamente, el problema de la inducción gira precisamente en torno a esta cuestión: ¿es posible una generalización siempre válida de los hallazgos? Desde el punto de vista de los defensores de la plausibilidad (Pólya, Jaynes, Cox, etc.), esta pregunta no se responde con un sí o un no, sino que se afirma que esta pregunta no tiene sentido porque los hallazgos siempre están contextualizados. La plausibilidad es, por tanto, la del intento de explicación E sobre otros intentos D, F o G.

Jaynes (2003, p.499) recoge esta idea cuando compara las estadísticas ortodoxas con las bayesianas. (véanse los capítulos 4.2.4 y 6.6 para una breve descripción de cada uno de ellos):

> „Fisher and Jeffreys, aware that all scientific knowledge has been obtained by inductive reasoning from observed facts, naturally enough denied the claim of Neyman that inference does not use induction, and of the philosopher Karl Popper that induction was impossible."

Y continúa (ibid., p.61),

> „...obviously, incomplete knowledge is the only working material a scientist has!" Además, da varios ejemplos para demostrar la difícil pero posible forma de justificación lógica de la inducción. Considera que es difícil especificar y aclarar los enunciados respectivos de forma que sean accesibles y puedan apoyarse en el análisis lógico.

El problema, entonces, es extraer la información disponible y eficaz del marco y aplicarla para resolver el problema, y luego evaluar las posibles explicaciones competidoras dada la información y los datos disponibles. Así, incluso con datos idénticos pero conocimientos previos diferentes (es decir, nivel de conocimiento sobre un problema), el proceso de razonamiento inductivo varía enormemente. Sin embargo, Jaynes (ibid., p.278) advierte que „[t]he fact that one person has far greater knowledge than another does not mean that they necessarily disagree; an idiot might guess the same truth that a scholar labored for years to discover." A continuación, sugiere que, para los mismos datos, una información previa diferente conduce a conclusiones diferentes (ibid., p. 279):

„The great variety of different conclusions that we have found from the same data makes it clear that there can be no such thing as a single universal inductive rule and, in view of the unlimited variety of different kinds of conceivable prior information, makes it seem dubious that there could exist even a classiffication of all inductive rules by some system of parameters."

Con respecto a los tratados filosóficos sobre la inducción (por ejemplo, en Carnap, 1973), Jaynes señala que con demasiada frecuencia se pierden en la lógica simbólica abstracta en lugar de tratar con ejemplos concretos y reales. Así, según Jaynes, los filósofos no entienden que las diferentes reglas inductivas corresponden simplemente a estados de información previa diferentes. Los distintos estados del conocimiento permiten, en primer lugar, nombrar los problemas y, en consecuencia, resolverlos, porque „there is no 'general inductive rule'" (ibíd.). No se trata de aplicar una única regla inductiva general, sino de aplicar reglas inductivas adecuadas al tema en cuestión en el contexto de la información previa. Dado que la estadística frecuentista ortodoxa (véase el capítulo 4) no considera la información previa (excepción, véase también Gelman & Carlin, 2014) ni la incluye en los cálculos, la inducción es aparentemente imposible desde este punto de vista. Más concretamente, hay que precisar que esto es imposible en la teoría de Neyman-Pearson (véase el capítulo 4.3.3), ya que el propio estadístico ortodoxo Fisher (véase más arriba) hablaba de inferencia inductiva y le concedía gran importancia.

Por cierto, no es la inducción lo que está bien o mal. En presencia de poca información, las conclusiones recurren a pocos datos y, por tanto, son menos precisas de lo deseado. La utilidad de la inducción en la ciencia, según Jaynes, no consiste en responder a la pregunta de qué predicciones son (deben ser) VERDADERAS, sino qué predicciones son las más plausibles dadas las hipótesis y la información disponibles. En un mundo de verdad relativa, la plausibilidad condicional es más importante que una verdad absoluta inalcanzable. La comparación de hipótesis contrapuestas puede ser fructífera aunque sepamos que algunas de ellas son falsas o sólo ligeramente plausibles. No obstante, en el uso hipotético de estas hipótesis podemos aprender algo sobre qué consecuencias prácticas (pueden) producirse y cómo cambian las condiciones, es decir, cómo se comportan los datos ante distintos modelos. Esto desplaza el problema. Por ahora se trata menos de un VERDADERO o un FALSO en el sentido de la validez de una lógica de decisión binaria. Ahora se trata más bien de comprender las relaciones causa-efecto para tomar decisiones basadas en la racionalidad, que esperemos sean prometedoras en la realidad.

Según Jaynes, la cognición por inducción no surge cuando crece la confianza en lo que ya se sabe, sino cuando se derrumba por inducción (Jaynes, 2003, p.326, cursiva en el original):

„... [T]he quickest path to discovery is to examine those situations where it appears most likely that induction from our present knowledge will fail. But those inferences must be our best inferences, which make full use of all the knowledge we have. One can always make inductive inferences that are wrong in a useless way, merely by ignoring cogent information.

Indeed, that is just what Popper did. His trying to interpret probability itself as expressing physical causation not only cripples the applications of probability theory ... (it would prevent us from getting about half of all conditional probabilities right because they express logical connections rather than causal physical ones), it leads one to conjure up imaginary causes while ignoring what was already known about the real physical causes at work. This can reduce our inferences to the level of pre-scientic, uneducated superstition even when we have good data."

Así pues, aprendemos algo de los acontecimientos observados precisamente cuando la inducción demuestra que las predicciones resultan erróneas o inadecuadas. Además, Jaynes advierte del peligro de confundir conexiones lógicas y causalidades físicas cuando se trabaja con probabilidades. En el mismo sentido se expresa Jeffreys (1931, p. 14):

„A common argument for induction is that induction has always worked in the past and therefore may be expected to hold in the future. It has been objected that this is itself an inductive argument and cannot be used in support of induction. What is hardly ever mentioned is that induction has often failed in the past and that progress in science is very largely the consequence of direct attention to instances where the inductive method has led to incorrect predictions."

Si resumimos las observaciones, apuntan a la defensa de una *cultura del error* radical para hacer posible un conocimiento nuevo y sustancial. Esto contrasta con la lógica científica de la publicación, en la que es evidente que *sólo merece la pena publicar aquello que pueda demostrar de algún modo resultados positivos (¡significativos!)*.

Así pues, la cuestión de la inducción se ha desplazado. No se trata de extraer una conclusión de generalización sobre todas las consideraciones posibles y concebibles (hipótesis, teorías, etc.) para producir o rechazar pruebas positivas, o como quiera que se llame esto. Más bien debemos tomar el camino contrario, es decir, formular siempre las afirmaciones en función de la información previa disponible y no desde la perspectiva de las infinitas posibilidades. Aquí es donde se encuentran Jaynes y Popper, que consideraban central el examen crítico de las teorías.

Dada la evidente probabilidad de errar en el razonamiento inductivo, parece casi increíble que la inducción pueda a veces conducir realmente a verdades empíricas. Por lo tanto, se pregunta Pólya (1954a), ¿debemos empezar analizando los casos en los que la inducción fracasa o examinar más de cerca los raros pero valiosos casos en los que la inducción resulta verdadera? En cualquier caso, necesitamos una cultura del error para obtener una comprensión más completa de los fenómenos de nuestro objeto de investigación concreto, anclada en la observación empírica. Se trata de establecer probabilidades relativas condicionales dada la información de que disponemos. Se trata del razonamiento bayesiano, que no se refiere a operaciones abstractas de lógica simbólica para aprender la inducción, sino a problemas concretos que conducen a conclusiones igualmente concretas.

De nuestra afirmación (véase el capítulo 1.1) de que en ciencia sólo existen verdades relativas, podemos deducir directamente algo sobre la inducción: Una afirmación como "No hay inducción" es en sí misma una forma de inducción, y no es válida cuando se aplica a sí misma. De lo contrario, se emitirían juicios sobre todo un sistema a partir de las afirmaciones de un subsistema o subsistema, lo cual no funciona (Whitehead & Russel, 1910). Si aplicamos esta afirmación a sí misma preguntando "¿Es cierto que no hay verdad (o inducción o lo que sea) en la ciencia?", rápidamente se hace evidente que, o bien nos estamos acercando a gran velocidad a la fase de pérdida de la realidad y quizá nos estemos volviendo un poco psicóticos, o bien hemos encontrado un verdadero problema que nos acompañará muy fielmente y durante mucho tiempo en la ciencia. Si nos alejamos de la realidad para adentrarnos en el mundo de la lógica pura, es más fácil vivir (por ejemplo, la inducción matemática), porque allí todo se trata en un espacio protector delimitado de la realidad. Pero incluso esta escapatoria tiene límites, como demuestra la conocida paradoja lógico-matemática "Todos los cretenses mienten" (Russell, 1908):

1. Heracles es cretense
2. Heracles dice: Todos los cretenses mienten.
3. ¿Es cierta la afirmación o miente Heracles?

En este punto de pérdida incipiente de la realidad, nos ayuda de nuevo el pensamiento bayesiano, en el que ya no nos hacemos la pregunta o preguntas anteriores, sino que, ante la información incompleta sobre el mundo, recopilamos la información de que disponemos y hacemos una especie de balance de caja:

1. ¿Qué información disponible habla en favor de la aplicabilidad de la inducción a un problema concreto?
2. ¿Qué información disponible habla específicamente en contra?
3. A continuación sacamos conclusiones, recopilamos nuevos datos, tenemos en cuenta los datos disponibles hasta el momento y repetir los pasos 1 a 3 hasta que podamos resolver el problema concreto, anclado en el contexto.

Ya no se trata de si la inducción es posible o funciona, sino de en qué condiciones puede aplicarse y con qué éxito o fracaso. En el pensamiento bayesiano, podemos asignar probabilidades a las afirmaciones 1 y 2 anteriores, que representan nuestro estado actual de información sobre el mundo. Nuestras afirmaciones derivadas de ellos son válidas a la vista de la información de que disponemos. Por supuesto, pueden cambiar y cambiarán. Si la información cambia debido a nuevos hallazgos (por ejemplo, datos empíricos), entonces cambian nuestras afirmaciones y todas las conclusiones derivadas de ellas. A continuación, las hipótesis en

liza pueden compararse de nuevo para ver qué posición está mejor respaldada por los datos empíricos o (por ejemplo, Gelman y Hill, 2007) si las hipótesis en liza pueden integrarse en un modelo más complejo. En principio, este proceso puede repetirse infinitas veces. Técnicamente, esto es posible gracias al teorema de Bayes (véase el capítulo 6.4). Presupone que nuestra información es precisa.

En sentido estricto, debemos considerar además que la propia información también es una fuente de entropía. Si tenemos motivos para categorizar cierta información como más precisa, ésta puede ponderarse de forma que tenga una mayor influencia. Si suponemos que la información es más o menos fiable, las probabilidades condicionales deberían estabilizarse a medida que aumenta la información (datos).

El físico Richard. T. Cox (1898-1991) en su obra fundamental sobre "The Algebra of Probable Inference" (Cox, 1961). Además de explicaciones detalladas de la lógica formal de la inferencia basadas en los trabajos de John Venn (1834-1923) y George Boole (1815-1864), se ocupa en particular de los cambios de probabilidades ante la inferencia inductiva, siempre coherente con el álgebra de Boole. La conjunción y la disyunción ocupan cada una una posición central (Cox, 1961, p.7):

> „Thus we see that
> $\sim (a \cdot b) = \sim a \vee \sim b$.
> From this equation and the equality of $\sim\sim a$ with a, there is derived a remarkable feature of Boolean algebra, which has no counterpart in ordinary algebra. This characteristic is a duality according to which the exchange of the signs, $\cdot$ and $\vee$, in any equation of propositions transforms the equation into another one equally valid. For example, exchanging the signs in this equation itself, we obtain
> $\sim (a \vee b) = \sim a \vee \cdot \sim b$,
> which is proved as follows:
> $a \vee b = \sim\sim a \vee \sim\sim b = \sim(\sim a \cdot \sim b)$."

En consecuencia, la inferencia axiomática se desarrolla a partir de las probabilidades condicionales, en las que a su vez se basa toda la estadística bayesiana. La coherencia de la inferencias se refiere a todas las ecuaciones y a las funciones contenidas en ellas. Las funciones pueden ser arbitrarias siempre que la ecuación global siga siendo coherente con el álgebra de Boole. La elección de la escala estadística que representa las probabilidades está sujeta a convenciones y no se deduce de las propias ecuaciones (ibíd., p.16).

> „Similarly, ifwe say that an inference is '95 per cent certain',we are saying that its probability is 95 on a scale on which certainty has the probability of 100. Usually it is convenient to represent certainty by 1 and, with this convention, the equation for the probability of the conjunctive inference is
> $i \cdot j|h = (j|h)\,(j|h \cdot i)$.
> [...] It is worth remarking, however, that other scales beside the ordinary one are consistent with this equation."

Cox demuestra la diferencia entre el razonamiento lógico y el mero cálculo de probabilidades con un conocido ejemplo de Laplace. Asumiendo la validez de la evidencia histórica de que se produjeron los últimos 1826213 amaneceres, Laplace calculó la probabilidad esperada del siguiente amanecer según la "regla de sucesión" como 1826214/1826215, es decir como ~99,999%. Sobre esto Cox comenta (ibid., p.89),

> „This calculation ignores the fact that, if one sunrise failed to occur as expected, this would, on any credible hypothesis, change the probability of the one expected to follow it."

Al final de su obra, Cox se ocupa específicamente del razonamiento inductivo. Lo define desde una perspectiva formal-lógica (ibid., p.91):

> „Inductive reasoning, when the term is used broadly, is any reasoning in which the verificationof one or more propositions is adduced as an argument for the truth, or at least the probability, of a proposition which implies them. For example, we see leaves moving and infer that the wind is blowing, or we hear the whistle of a locomotive and infer that a train is coming.
>
> The argument depends on the equality of the two expression for the probability of a conjunctive inference. Let g be a proposition which, on the hypothesis h, implies another proposition, i. Equating the two expressions for g i | h, we have

(g | h * i) (j | h) = (i | h * g) (g | h);

whence

$$\frac{g\,|\,h\cdot i}{i\,|\,h\cdot g} = \frac{g\,|\,h}{i\,|\,h}.$$

To say that g implies i is to say that i | h * g = 1 and thus

$$g\,|\,h\cdot i = \frac{g\,|\,h}{i\,|\,h}. \quad (18.1)$$

[...] Eq. (18.1) shows that the verification of any proposition i increases the proability of every proposition g which implies it."

Las explicaciones de Cox demuestran la lógica básica de asignar una probabilidad a los sucesos condicionales basándose en la ocurrencia (probabilística) de sus condiciones. El teorema de Bayes no significa otra cosa, aunque Cox no lo diga aquí. Así, una conclusión inductiva precisa y coherente con el álgebra de Boole puede expresarse mediante el teorema de Bayes. En Jaynes (2003, cap. 2) se puede encontrar una derivación y discusión detalladas. Para la ecuación de Cox 18.1 anterior, cuanto menor sea *i* | *h*, es decir, la probabilidad a priori de *i*, mayor será el cociente *(g* | *h* · *i)/(g* | *h)*, es decir, el factor cuya prueba positiva aumenta la probabilidad *g*. En lenguaje cotidiano, esto significa que si un suceso condicional g sólo tiene una probabilidad previa baja debido a sus condiciones improbables, entonces la probabilidad condicional o expectativa de que ocurra este suceso cambia tanto más cuando sus condiciones y, en consecuencia, el suceso g ocurren realmente. Esto provoca un cambio significativo en la expectativa respectiva, es decir, en la probabilidad del suceso g. En la inducción se hace hincapié en un conjunto de en principio infinitas secuencias de condiciones y no en una condición singular que guarde una relación lógica con el acontecimiento de interés. La inferencia por inducción como tal es de naturaleza probabilística dentro de un contexto finito concreto (ibíd., p. 93).

„The ensemble which is made the subject of an induction is ordinarily unlimited in the number of instances. The argument is aimed at establishing a universal principle, valid under given circumstances no matter how many times they are encountered or produced. Certainty is hardly to be expected in such an argument, for it would be surprising if a principle could be proved valid in an infinite number of instances by being verified in a finite number. In some cases, however, certainty is approximated when the number of verified instances is very large."

Cox aborda así directamente la crítica de Hume a la inferencia por inducción, ya que suponía que el conjunto de condiciones no cambiaría mucho de un caso a otro. Cox responde a esto (ibid., p.94),

„The instances differ more among themselves, however, than is implied in Hume's question. They must differ in some resepect in order to be distinguishable one from another and they may differ with respect to any characteristic except that by which the ensemble is defined. Specifically, with respect to the characteristic in question in the induction, the instances are not know to be alike until their likeness is verified by observation. This verification provides a ground for inference which was not present before."

Cox ve la crítica de Hume más bien como una indicación de que estaba corrigiendo la opinión errónea de muchos, que suponía que en la inducción no sólo había una aproximación gradual a la certeza completa, sino que esta certeza total también era alcanzable en la realidad. Este no es el caso. Por otra parte, la inducción sigue sus propias leyes que, según Cox, no pueden derivarse de las de la deducción. Cox entiende la crítica de Hume de que la inducción, además, no es racional en el sentido de que Hume equiparaba la racionalidad con la certeza total y la deducción. En este sentido, para Hume el razonamiento se inscribe siempre en la tradición deductiva.

La forma de inducción representada aquí sobre la base de la información contextual previa explica ahora cuándo decisiones son *sabias o no* (Studer, 1996b; Jaynes, 1986a, p.2).

> "Long before studying mathematics we have all learned, necessarily, how to deal with such problems intuitively, by a kind of plausible reasoning where we lack the information needed to do the formal deductive reasoning of the logic textbooks. In the real world, some kind of extension of formal logic is needed. [...]"
>
> From the earliest times this process of plausible reasoning preceding decisions has been recognized. Herodotus, in about 500 BC, discusses the policy decisions of the Persian kings. He notes that a decision was wise, even though it led to disastrous consequences, if the evidence at hand indicated it as the best one to make; and that a decision was foolish, even though it led to the happiest possible consequences, if it was unreasonable to expect those consequences."

Tabla 2.5: *La conclusión deductiva*

	Abstracción formal	**Ejemplo**
Premisa principal (ley, regla)	Si A, entonces B.	Si llueve, entonces la carretera está mojada.
Premisa secundaria (caso)	Dado A,	llueve.
Conclusión (resultado)	Entonces B.	La carretera delante de mi casa está mojada.

Así pues, el *sentido común*, que ya menciona Heródoto en torno al año 500 a.C. y es probable que sea bastante más antiguo, demuestra su sabiduría si se actúa de acuerdo con la información disponible. Si no se hace así, una decisión se considera insensata, independientemente del resultado real. Esto significa que la expectativa sobre el curso de acción posterior se desarrolla a partir de la información disponible y este proceso constituye la base de la evaluación de una decisión denominada razonable. Por lo tanto, es significativo que la sabiduría de la toma de decisiones se considere aisladamente del resultado real. Se trata de elegir entre cursos de acción alternativos en función de sus respectivas probabilidades esperadas. Una decisión puede ser sabia y aún así conducir a un resultado desastroso y, a la inversa, una decisión puede ser tonta y conducir a un resultado exitoso. En ambos casos, sin embargo, esto no era evidente debido a las expectativas (es decir, información o datos conocidos). Esto muestra claramente la diferencia entre ciencia y clarividencia. La ciencia se desarrolla según reglas definidas. La clarividencia corresponde a un enfoque intuitivo basado en percepciones sensoriales. Esta aparente discrepancia refleja simplemente la calidad de la información disponible y no el proceso de toma de decisiones en sí. Una evaluación del proceso de toma de decisiones es posible sin conocimiento del resultado, es decir, únicamente sobre la base de la combinación de la información disponible y sus conclusiones respecto a una expectativa. Concluimos el tema del razonamiento inductivo con otra cita de Cox que expresa todo esto junto (Cox, 1961, p.96).

> „If there be any possibility that the course of nature is uniform and that the past may be some rule for the future, all experience becomes useful and can give support to some inference."

2.2 La conclusión deductiva

En cambio, podemos estar absolutamente seguros cuando concluimos deductivamente. Los orígenes se remontan a la enseñanza de Aristóteles (384-322 a.C.) sobre los *silogismos*. Observó que tales inferencias tienen una estructura formal (véase la Tab. 2.5, incluido el ejemplo de la lluvia). Veamos de nuevo la deducción con el ejemplo inicial de la lluvia y la carretera. La conclusión de una regla dada y un caso dado

es absolutamente cierta o VERDADERA debido al componente "entonces" ("la carretera está mojada") de la regla - pero es muy problemático científicamente.

La conclusión no nos aporta nuevos conocimientos, sino que sólo confirma lo que ya conocemos como ley. O dicho de otro modo: en la deducción concluimos desde lo general (ley, norma) a través de la ocurrencia del caso (premisa secundaria) hasta lo particular (resultado). Veremos, sin embargo, que en el trabajo de investigación la interacción de la inducción y la deducción es indispensable y la conclusión de lo general a lo particular puede sin duda llevar más lejos a los investigadores. Por ejemplo, el "método de comparación constante" del enfoque de la "teoría anclada" (Glaser & Strauss, 1967) se basa en la interacción del razonamiento inductivo y deductivo en el análisis de datos cualitativos. Sin embargo, una conclusión deductiva formalmente correcta no es necesariamente VERDADERA en su contenido, es decir, si la proposición o enunciado en el que se formuló la ley es falso. A partir de una regla "cuando llueve, la calle se ilumina" y del enunciado del caso concreto "llueve", podríamos deducir lógicamente de forma correcta que ahora se ilumina la calle, un caso que probablemente no se dé en la realidad. La deducción permite entonces saber exactamente cuándo algo no tiene lugar, es decir, cuándo una regla no se confirma empíricamente. Esto requiere una revisión o incluso una redefinición de la norma. Eso es (también) una comprensión.

Tabla 2.6: *Silogismos modus ponens y modus tollens*

	Modus ponens		**Modus tollens**	
	Abstracción formal	Ejemplo	Abstracción formal	Ejemplo
Premisa principal (ley, regla)	Si A, entonces B.	Si llueve, entonces la carretera está mojada.	Si A, entonces B.	Si llueve, entonces la carretera está mojada.
Premisa secundaria (caso)	A es VERDADERO.	Lluvia.	B es FALSO.	La carretera está seca.
Conclusión (resultado)	B es VERDADERO.	La calle está mojada.	A es FALSO.	No lluvia.

Tengamos en cuenta que una premisa es una proposición, es decir, una afirmación inequívoca. Puede ser VERDADERO o falso en el marco de un contexto determinado (situación problemática concreta) o puede resultar VERDADERO o falso. En caso de información incompleta o inexacta, es posible que no podamos deducir conclusiones con certeza. Esta es probablemente la norma, al menos en las ciencias sociales, independientemente del contexto.

Las inferencias o silogismos más conocidos son el modus ponens (del latín ponere: poner) y el modus tollens (del latín tollere: negar), como se muestra en el cuadro 2.6.

La aplicación de la deducción en el contexto de la investigación parece relativamente sencilla a primera vista. Por ejemplo, existe una teoría general que explica nuestro ejemplo anterior de alumnos poco dispuestos a esforzarse, es decir, por qué se abstienen de realizar esfuerzos de aprendizaje. A partir de esta teoría general, se puede generar una intervención pedagógica para el caso individual de un alumno en riesgo de fracaso escolar. Esta intervención se deduce, es decir, se deriva, de la teoría general sobre las causas de la renuncia al esfuerzo en los alumnos y las posibilidades de prevención relacionadas. Para ello es necesario que esta teoría general contenga enunciados que expliquen el comportamiento de los alumnos y permitan posteriormente deducir conclusiones para la práctica pedagógica. Por supuesto, es exigente derivar medidas concretas para los individuos a partir de una teoría general, es decir, medidas que correspondan realmente a la conclusión deductiva y estén ancladas en la teoría. Debido a la pretensión general de amplio poder explicativo y aplicabilidad, los fundamentos teóricos de los estudios empíricos actuales nunca alcanzan realmente el estatus de una teoría general, sino que, en el mejor de los casos, son en su mayoría

planteamientos teóricos. En la práctica, se trata de fragmentos que a menudo se parecen más a una alfombra de retazos.

Es más exigente derivar de una teoría hipótesis comprobables empíricamente. Eso es mucho más difícil que formular hipótesis estadísticas que puedan distinguirse de ella. La forma clásica de *explicación deductivo-nomológica* como distinción de la explicación cotidiana (no científica) ofrece el *esquema Hempel-Oppenheim* (Hempel & Oppenheim, 1948). Describe la argumentación lógicamente correcta que consiste en el enunciado científico generalmente válido de la ley y la observación empírica (Explanans, del latín explanare: interpretar, explicar), así como la cosa a explicar (Explanandum, gerundio de explanare), que se deduce. En resumen, lo que se explica se deriva de lo que está de lo que está explanando (véase la tabla 2.7). Aunque el explanandum resulta necesariamente del explanans, éste debe tener un contenido empírico. Esto significa que debe ser falsable, es decir, que en principio debe ser posible averiguar lo contrario.

Tabla 2.7: *El esquema Hempel-Oppenheim*

	Abstracción formal	**Ejemplo**
Explanans - lo que explica algo	Enunciado de ley + Observación empírica	Si llueve, entonces la carretera está mojada. + Lluvia
Explanandum - lo que está explicando algo	Conclusión lógica	La carretera está mojada.

Dado que en las ciencias sociales complejas, a diferencia de las ciencias naturales simples (por ejemplo, la física), no existen enunciados de ley de validez general, normalmente hay que trabajar con hipótesis y fragmentos teóricos para poder hacer predicciones. Éstas, a su vez, pueden comprobarse mediante la observación empírica a partir de explica-ciones y predicciones teóricas. Asignar un alumno a la formación por dislexia y el posterior seguimiento del progreso del aprendizaje basado en una prueba diagnóstica de la dislexia correspondería a un modelo explicativo deduc-tivo-nomológico debilitado de esta manera (Groeben, 1986) (véase la tabla 2.8).

Tab. 2.8: *Explicación deductiva-nomológica*

	Ejemplo
Explanans (Enunciación de ley)	La dislexia debe presuponerse a partir de 20 errores en el test.
Observación empírica (Condición de contorno)	Xana cometió 20 errores.
Explanandum (Conclusión lógica)	Xana padece dislexia.

Por supuesto, no es una ley que se aplique a todas las personas en todas las circunstancias. En cambio, una manzana cae inevitablemente de arriba abajo en cualquier momento del día y del año, a menos que, por ejemplo, un tornado contrarreste la fuerza de la gravedad. Por supuesto, la manzana también atrae a la Tierra y, estrictamente hablando, ambas se aproximan, salvo que la atracción gravitatoria de la manzana es despreciable. El término ley induce a error en las ciencias sociales, a menos que intervengan condiciones biológicas o físicas (sistema nervioso, actividad cerebral, memoria, etc.). Se trata más bien de declaraciones de normas contextualizadas. La calidad diagnóstica de una prueba de rendimiento, por ejemplo, puede ser alta o baja. Sin embargo, nunca alcanza un nivel determinista. E incluso con recortes, las predicciones nunca son tan exactas como las ecuaciones de Maxwell o los cálculos de Newton sobre las órbitas planetarias, que tuvieron lugar antes del desarrollo de los ordenadores y los instrumentos de medición precisos. Demasiadas

influencias y fluctuaciones afectan al resultado de la prueba. Las pruebas siempre se calibran para muestras grandes y las afirmaciones sobre casos individuales son afirmaciones de tendencia que no tienen carácter determinista, sino probabilístico. Siempre hay circunstancias acompañantes que influyen en el resultado de la prueba. Por ejemplo, un niño puede desarrollar ansiedad durante una prueba de este tipo y quedarse por debajo de sus propias capacidades debido a la excitación, o simplemente puede estar teniendo un mal día, estar cansado, etc. Además, la famosa "rotura del lápiz" puede costarle al niño valiosos puntos en la prueba de velocidad (es decir, a tiempo). Es posible que el criterio de 20 errores utilizado en el ejemplo anterior se haya fijado tras muchas reflexiones y pruebas. Sin embargo, no es válido para siempre, al igual que los tests de inteligencia tienen que recalibrarse una y otra vez. ¿Qué tal, por ejemplo, 19 o 21 en lugar de 20 errores? ¿Podemos afirmar o negar el diagnóstico de dislexia con la máxima certeza? Sólo si añadimos más fuentes de información, un juicio incierto se convierte en un diagnóstico plausible que permite realizar intervenciones (por ejemplo, entrenamiento en lectoescritura). No otra cosa aparece más tarde en la estadística clásica (véase el capítulo 4.3.8) cuando se trata de la cuestión de la interpretación en los límites de las barreras de signi?cancia, es decir, cuando se aborda la cuestión de hasta qué punto estamos seguros de haber encontrado algo significativo frente a algo insignificante (véase también Gelman y Stern, 2006). Sin embargo, todo esto no cambia la forma de razonar deductivamente del resultado de la prueba a la presencia (o no) de dislexia. El proceso deductivo sigue siendo el mismo. Sólo cambia la realidad.

La lógica abstracta puede ser una herramienta útil en el análisis de datos. En relación con la lógica abstracta pura, la R permite realizar directamente consultas lógicas (ptI_sciencetheory_logic.r)[1]:

```
> TRUE & TRUE
[1] TRUE
> TRUE & FALSE
[1] FALSE
> FALSE & FALSE
[1] FALSE
> FALSE | FALSE
[1] FALSE
```

así como en forma compleja

```
> (TRUE || FALSE) & (FALSE & TRUE)
[1] FALSE
```

y por supuesto relacionados con objetos que pueden tomar los valores TRUE o FALSE:

```
> x <- 1:10
> x >= 5 & x <10
[1] FALSE FALSE FALSE FALSE TRUE TRUE TRUE TRUE TRUE FALSE
> x[x >= 5 & x <10]
[1] 5 6 7 8 9
> x >= 5 && x <10
[1] FALSE
```

2.3 La conclusión abductiva

[1] Los scripts R que contienen el código R que coincide con los estudios de caso se enumeran entre paréntesis con sus nombres la primera vez que se utilizan. Los scripts de R son más extensos en código de lo que se imprime aquí. Sólo se enumeran el código R pertinente y el resultado resultante para ayudar a comprender los estudios de casos. Para más detalles, se ejecuta el código R paso a paso.

Según Charles S. Peirce, la abducción se considera una lógica de la inferencia por derecho propio (Reichertz, 2000). Goza de gran popularidad en las ciencias sociales de orientación cualitativa. En filosofía, la abducción es la conclusión formal del resultado (conclusión) y la regla (premisa principal) al caso dado (premisa secundaria) (Schmidt, 1991).

Según Peirce, el punto de partida y la base de la abducción es la percepción (Peirce, 1965, CP 5.171 o CP 5.189) sobre la base de la experiencia previa y las correspondientes expectativas formadas acerca de los acontecimientos observables:

> „It must be remembered that abduction, although it is very little hampered by logical rules, nevertheless is logical inference, asserting its conclusion only problematically or conjecturally, it is true, but nevertheless having aperfectly definite logical form.
>
> Long before I first classed abduction as an inference it was recognized by logicians that the operation of adopting an explanatory hypothesis — which is just what abduction is — was subject to certain conditions. Namely, the hypothesis cannot be admitted, even as a hypothesis, unless it be supposed that it would account for the facts or some of them. The form of inference, therefore, is this:
>
> The surprising fact, C, is observed;
>
> But if A were true, C would be a matter of course,
>
> Hence, there is reason to suspect that A is true.
>
> Thus, A cannot be abductively inferred, or if you prefer the expression, cannot be abductively conjectured until its entire content is already present in the premiss, 'If A were true, C would be a matter of course.'"

Para Peirce, la abducción es el proceso mediante el cual se desarrollan hipótesis explicativas (ibid., CP 5.171). Estos preparan el camino para la inducción. Por ejemplo, permite el razonamiento heurístico y la experimentación para reunir, a partir de casos individuales y utilizando hipótesis ad hoc y reglas formuladas heurísticamente, el material de razonamiento con el que pueden formularse reglas y leyes más generales hacia una conclusión inductiva correspondiente. A continuación, puede comprobarse deductivamente (nomológicamente), etc.

Tabla 2.9: *La conclusión abductiva*

Abstracción formal	**Ejemplo**	
Conclusión (resultado)	Así que B.	Hans no se esfuerza y por eso no aprueba el examen.
Regla (generación)	Si A, entonces B.	Si haces un esfuerzo y fracasas, entonces se considera que eres sin talento.
Norma alternativa (Regla de negación)	Así que A.	Si uno *no* se esfuerza y fracasa, entonces *no* se le considera sin talento. Hans no quiere ser considerado estúpido. Hay otros niños de la clase que no quieren ser considerados estúpidos.

Tomemos de nuevo nuestro ejemplo empírico. Hay un grupo de investigación pedagógica que se enfrenta al problema de no poder sacar conclusiones sobre su caso problemático ni deductiva ni inductivamente. La cuestión es que los niños no se esfuerzan en los exámenes escolares en determinadas condiciones, aún poco claras y, en consecuencia, no aprueban los exámenes. En este punto, los investigadores no tienen ni hipótesis para las que puedan desarrollar un diseño de prueba empírica, ni una teoría sólida con la que deducir intervenciones educativas. Más bien se han limitado a buscar una "ley" como

explicación plausible para explicarles el caso. Para ello, necesitan observaciones y pensamientos creativos. Más concretamente, los investigadores han observado un caso interesante en el campo de acción "clase escolar" y, sorprendentemente, no pueden explicarlo con las reglas que conocen. Esto significa que al principio ya no es necesario el razonamiento lógico, sino la intuición y la creatividad para encontrar una posible explicación a la observación. Los miembros del grupo deben idear una conexión – construir una regla – que pueda explicar la observación (véase el cuadro 2.9). Este es el resultado de una conclusión llamada *abducción*. Así, en este procedimiento, primero tenemos una conclusión o resultado, la percepción sorprendente. A continuación, se generan las reglas para aplicarlas a un caso (nuevo) y comprobarlas heurísticamente. Este proceso corresponde en principio a una aplicación mixta de intuición y corazonada, inducción y deducción, al siguiente caso individual. Coinciden de manera no formalizada.

Lo que se observa es el imprevisto de que Hans no se esfuerce en un examen. Reflexionar sobre las posibles explicaciones de la falta de esfuerzo lleva a establecer una regla ad hoc. Aquí se afirma que la falta subjetiva de esfuerzo protege contra la atribución estable de falta de capacidad. Esto podría dar lugar a más suposiciones. Si pensamos que la regla es plausible, nos fijamos en otros niños de la misma clase que quizá tampoco quieran ser considerados estúpidos. Estas inferencias, llamadas "rayo abductivo" por Peirce (1965, CP 5.181), describen el proceso de observación (inesperada) y generación de reglas hipotéticas y aplicación de las mismas al siguiente caso individual. La norma parece bastante plausible en este caso y podríamos preguntarnos además cómo es posible que Hans y muchos otros alumnos prefieran parecer vagos a estúpidos y por eso renuncian al esfuerzo escolar. Se trata de una pregunta de investigación para la que se podría desarrollar un diseño de investigación y buscar respuestas. A continuación, examinaríamos la autoestima, el clima de clase, la relación profesor-alumno, el nivel de exigencia y (por ejemplo, mediante un test de CI) el posible nivel de capacidad de Hans y de otros alumnos en diferentes contextos (escolar, extra-escolar). A la inversa, también pueden derivarse de ello intervenciones de acción concretas, dirigidas menos al rendimiento cognitivo que a la motivación o a los componentes sociales. Esto es exactamente lo que hizo Jopt (1978) basándose en la teoría de la atribución de causas de éxito y fracaso. Llegó a la conclusión de que la aparente "pereza", es decir, la evitación de esfuerzo, puede interpretarse como una estrategia "plausible" y comprensible de los alumnos. La evitación de esfuerzo tiene como objetivo evitar una amenaza para el auto-concepto explicando los fracasos con una renuncia deliberada al esfuerzo personal, evitando así la cuestión de la atribución interna de la capacidad en primer lugar. Por cierto, esto protege contra la sensación de pérdida de control, ya que la evitación de esfuerzo vuelve a ser intencionada. Las consecuencias en la realidad son, sin embargo, una caída masiva del rendimiento, que puede desembocar rápidamente en un ciclo.

Notemos que el resultado de la abducción es la conclusión lógica de algo especial (caso individual) a otra cosa especial (a saber, el siguiente caso individual) mediante la generación de una regla hipotética correspondiente y mediadora. Si se confirma empíricamente, esta norma hipotética puede convertirse en una norma justificada y aceptada con un ámbito de aplicación definido, y por lo tanto ocupar un lugar en el canon científico. Del mismo modo, el proceso también puede conducir directamente a la falsificación. La abducción no distingue estos casos: confirmación positiva frente a prueba de falta de eficacia.

La base de teorías potencialmente útiles desde el punto de vista científico basadas en observaciones cotidianas ha sido retomado por diversos autores (Kelly, 1955; Groeben, 1986). El *programa de investigación "Teorías subjetivas"* (Groeben y Scheele, 1977), por ejemplo, intenta formalizar precisamente este proceso y combinar la comprensión o descripción por un lado y la observación o explicación por otro (integración monismo-dualismo).

Bauer (2000, véase también Stangl, s.f.) cita un ejemplo de razonamiento abductivo políticamente incorrecto y con intenciones poco serias:

> Man: „Hi there, new neighbor, it sure is a mighty nice day to be moving."
> Neighbor: „Yes, it is and people around here seem extremely friendly."
> Man: „So what is you do for a living?"
> Neighbor: „I am a professor at the University, I teach abductive reasoning."
> Man: „Abductive reasoning, what is that?"
> Neighbor: „Let me give you an example. I see you have a dog house out back. By that I abduce that you have a dog."

Man: „That is right.“
Neighbor: „The fact you have a dog, leads me to abduce that you have a family.“
Man: „Right again.“
Neighbor: „Since you have a family I abduce that you have a wife.“
Man: „Correct.“
Neighbor: „And since you have a wife I can abduce that you are heterosexual.“
Man: „Yup.“
Neighbor: „That is abductive reasoning.“
Man: „Cool.“
Later that same day...
Man: „Hey I was talking to that new guy who moved in next door.“
Neighbor 2: „Is he a nice guy?“
Man: „Yes, and he has an interesting job.“
Neighbor 2: „Oh, yeah what does he do?“
Man: „He is a professor of abductive reasoning at the University.“
Neighbor 2: „Abductive reasoning, what is that?“
Man: „Let me give you an example. Do you have a dog house?“
Neighbor 2: „No.“
Man: „Fag.“

La abducción se considera un proceso de descubrimiento heurístico y exploratorio para extraer estructuras, ideas, principios, reglas, etc. a partir de datos de cualquier tipo. Podemos hacerlo observando, analizando textos, mirando gráficos de datos, o simplemente observando, participando y experimentando, cerrando los ojos y dejándonos llevar por la imaginación. Esto haría legítimo preguntarse si la abducción no tiene también partes de la clarividencia antes mencionada, si la intuición tiene ese papel. La abducción combina la acción guiada por reglas con el acceso individual intuitivo al mundo. En este sentido, la abducción es una forma menos formalizada de razonar, lo que beneficia a la creatividad. Asimismo, forma parte de nuestra práctica cotidiana y, como muestra el ejemplo anterior, es una enorme fuente de estereotipos cuya validez en la realidad no comprobamos necesariamente ni mucho menos de forma fiable. Antes de inducir, seguro que – al menos en nuestra mente – ya hemos sacado varias conclusiones abductivas y hemos vuelto a rechazar casi otras tantas.

2.4 ¿Se aplica la trinidad al razonamiento científico?

Las explicaciones anteriores corresponden al modelo estándar de razonamiento lógico y plausible. Si ahora revisamos cómo se conectan la inducción, la abducción y la deducción, las incoherencias se hacen evidentes. Éstas se refieren principalmente a la cuestión de la relación concreta de la inducción y la deducción y, a continuación, a si la inducción, la deducción y la abducción se solapan o si, de hecho, son tres procesos claramente separados, cada uno de los cuales justifica entonces tres términos diferentes. Consideramos que la inducción y la deducción deben modelarse de tal manera que emerja clara y distintamente una relación dialéctica subyacente. Esto corresponde a una lógica dual clásica, como la que encontramos en el principio del yin y el yang. Y en lo que respecta a la abducción, se solapa tanto con la inducción como con la deducción. Esto requiere una división y fusión parcial de los tres conceptos. Ambas afirmaciones se aclaran a continuación.

La *deducción* como conclusión de lo general a lo particular está eficientemente formulada mediante la aplicación de la lógica formal y no es un problema, sino siempre una tarea. Esto no tiene nada de malo.

La *inducción* como razonamiento de lo particular a lo general es, por así decirlo, la inversión del razonamiento deductivo de lo general a lo particular. Tiene de principio un carácter abierto y, por tanto, es un problema y no una tarea. Las críticas a la inducción se basan en la afirmación de que no puede ser analítica ni ampliar el contenido. Estas críticas demuestran, sobre todo, que la filosofía no está muy cerca de la realidad. De hecho, si la inducción se contempla desde la perspectiva de la psicología, son esos procesos cognitivos los que agregan información para convertirla en conocimiento (Johnson-Laird, 1993). Esto no significa otra cosa que los procesos inductivos comprenden el conglo-merado de todos aquellos procesos

cognitivos que intentan dar sentido al mundo y crear una regla generalizada efciente de y para acontecimientos concretos a partir de la abundancia de infinita información individual. Evolutivamente, esto tiene sentido y puede considerarse equivalente a la heurística, que son mini-inducciones casi ad hoc. Obviamente, su aplicación ha aportado una ventaja evolutiva al poder afrontar mejor el mundo cambiante mediante la identificación de regularidades y, en consecuencia, una actuación rápida. Jaynes y Pólya no expresan otra cosa en su obra cuando hacen siempre hincapié en el razonamiento plausible (inductivo) contextualizado para dirigir el foco hacia un número finito de explicaciones posibles, lejos de la idea no comprensible de infinitas y absurdas explicaciones teóricas, que general-mente deben incluir ovnis, extraterrestres, deidades, seres invisibles y, sin olvidar, el calendario maya y el apocalipsis. Sin embargo, si el empirismo sugiriera algo de estas áreas de apariencia esotérica como significativo y fuera capaz de explicarlo teóricamente, estos conceptos inusuales y no cotidianos tendrían que ser investigados inmediatamente como potencialmente plausibles y científicos. Una idea de cómo una empresa tan delicada – delicada porque uno se expone rápidamente al ridículo en los círculos profesionales – puede parecer seria y científica y qué consecuencias tiene la analiza el físico e investigador de ovnis Illobrand von Ludwiger (2015).

Lo que sigue sin resolverse es la relación entre inducción y deducción. En la tradición aristotélica, ambas formas de inferencia no suelen considerarse dialécticamente, sino desde el punto de vista de la lógica silogística, que, al fin y al cabo, se utiliza para aportar pruebas. Sin embargo, cuando se genera una regla, esto no corresponde a una prueba, sino a una creación creativa para ver conexiones donde antes sólo se veía información incoherente. La inducción no es una forma de razonar, sino de crear algo nuevo (por ejemplo, mediante el reconocimiento de patrones, la comparación de casos, la búsqueda de regularidades y repeticiones, la identificación de desencadenantes de procesos, su categorización y denominación, la activación de procesos inconscientes, etc.). En este sentido, creemos que la tradición aristotélica de pensamiento es inadecuada para modelar con precisión la relación entre inducción y deducción. Y aquí es precisamente donde radica el error de pensamiento de Popper, que pretendía haber resuelto por completo el problema de la inducción eliminando la inducción y convirtiendo de golpe todos los procesos en exclusivamente deductivos por naturaleza. Parece mucho más sensato situar ambos procesos en una relación dialéctica socrática. Uno no puede existir sin el otro, pero son opuestos entre sí. Ambas se relacionan entre sí de un modo comparable a tesis y antítesis, sin que tenga sentido, por ejemplo, asignar la tesis a la deducción y la antítesis a la inducción, o viceversa. Sin embargo, juntos – sintéticamente – conducen a la posibilidad de agregar conocimientos a partir de la información, de reconstruir el significado y, posteriormente, de aplicar y probar empíricamente este conocimiento y este significado. Desde un punto de vista dialéctico, los términos deducción e inducción parecen ser meras puntuaciones arbitrarias (Bateson, 1985) del infinito flujo natural de percepciones, reconstrucciones de sentidos y sus aplicaciones o pruebas en la realidad, es decir, la vida. Para la in-vestigación empírica, la lógica formalmente abstracta es interesante, pero no necesariamente determinante, ya que repre-senta un ideal inalcanzable. Lo que es absolutamente necesario, sin embargo, es un razonamiento metodológicamente controlado y coherente sobre el caso concreto y a través de los casos, para luego pasar de nuevo al otro caso concreto o a un caso completamente nuevo.

La *abducción*, a su vez, se sitúa entre la inducción y la deducción y, desde nuestro punto de vista, se solapa con las otras dos formas de inferencia. Por un lado, se genera una regla basada en el caso individual (empirismo) y esto refleja el proceso básico de abstracción inductiva de lo particular a lo general. A la inversa, esta regla abstracta recién generada se aplica al siguiente caso individual, lo que refleja de nuevo la inferencia deductiva de lo general a lo particular. Visto desde fuera, ambos procesos coinciden, de modo que puede parecer que un caso individual se infiere al siguiente caso individual mediante un *rayo abductivo*. Esto parece distinguir la abducción de la inducción. Sin embargo, es mucho más complejo de lo descrito. En nuestra opinión, la abstracción es esencial en el razonamiento abductivo debido a la necesidad de generación de reglas y no sólo se concluye en el nivel concreto de caso a caso, sino de caso a regla – y de regla a caso. Siempre hay que tener en cuenta la abstracción temporal. A la inversa, la regla generada ad hoc sólo puede inferirse al siguiente caso individual si se utiliza un conjunto mínimo de reglas deductivas para establecer un modelo predictivo basado en la regla, de modo que se pruebe provisionalmente en el caso individual, se revise, se abandone, etc. Así pues, en cuanto se introduce la abstracción y el nivel de las reglas, que a su vez se aplican al caso, los aspectos originales de la inducción o la deducción ya han sido solapados por los

procesos abductivos. De este modo desaparecen los criterios de distinción más importantes frente a la inducción y la deducción. Así pues, ya no es posible una separación clara entre abducción, deducción e inducción, y habría que preguntarse si la abducción no describe "simplemente" el proceso iterativo de generación de reglas que tiene lugar realmente en la gente corriente, que se basa en la percepción intuitiva y bastante gestáltica, la generalización creativa, la comprobación, la adaptación y la revisión, etc. Esto es coherente con la descripción de la abducción y la deducción. Esto es coherente con la visión de Peirce descrita anteriormente, que entendía el conocimiento no como un ?xen estado o hecho objetivo, sino como un proceso que siempre incluía la percepción – y, por tanto, tenía en cuenta el elemento concreto y contextualmente situado – , visión que presumiblemente también podrían sostener los representantes de las teorías de la verosimilitud y la estadística bayesiana. Y aunque el conocimiento esté formalmente escrito y, por tanto, simplificado, es el resultado de un proceso y no el proceso en sí. El proceso en sí parece mucho más caótico y creativo o intuitivo. Como se explica en el capítulo 2.3, la abducción se caracteriza precisamente por estos elementos de creatividad e intuición. En sentido estricto, esto no es en absoluto necesario para la deducción, pero sí lo es mucho para la inducción, cuando es necesario ?encontrar conexiones inesperadas entre experiencias individuales que antes no existían. Y formular a continuación una regla más o menos válida en general a partir de estas referencias es un verdadero acto creativo de creación.

Ahora podemos preguntarnos: ¿es (todavía) necesaria esta división tripartita? La respuesta podría ser: en realidad, ya no. Más bien, lo que se necesita es una adaptación y reestructuración de las formas de concluir sin descartar los detalles elaborados. La siguiente subdivisión de las secuencias tiene fines puramente analíticos. En la práctica, las fases van de la mano. Un análisis exhaustivo del tema y de la ejecución a nivel de detalle llenaría sin duda una obra de varios volúmenes. Por lo tanto, aquí sólo esbozaremos la idea básica:

1. Generación de reglas y frases basadas en percepciones (por ejemplo, observación, memoria, etc.), que se ponen en un nuevo contexto mediante diversos procesos cognitivos (por ejemplo, comparación, formación de categorías, imaginación y simulación interna/experimento mental o intuición, etc.). Contienen un elemento creativo o intuitivo explícito, que no puede modelarse algorítmicamente (es decir, se trata de un problema no trivial). Pero en principio puede resolver tanto problemas triviales como no triviales. Por tanto, los humanos pueden resolver problemas de investigación y encontrar reglas, pero no necesariamente mediante la aplicación de un algoritmo, sino a través de la creatividad y la intuición humanas. Esto es lo que nos distingue (¿actualmente?) de la IA.
2. Aplicación del conjunto de reglas generado al caso concreto y examen empírico crítico del mismo.
3. Revisión, rechazo, ampliación, etc. del conjunto de normas y recogida de nueva información (percepciones).

Estas tres fases reflejan el ciclo de investigación en espiral esbozado al principio (véase la Fig. 2.1, p. 8). Por las razones antes mencionadas, proponemos combinar la abducción con la inducción, mantener la deducción sin cambios, cambiar el nombre del "nuevo" conglomerado de inducción y abducción, e incluir explícitamente la creatividad o intuición como parte del razonamiento y modelarla como una herramienta inesperada e incontrolable pero significativa para el progreso. Esto añade mucha más incertidumbre al modelo global, que por definición no puede modelizarse de forma determinista. Por buenas razones, porque en este ámbito de la ciencia se trata de crear algo nuevo y esto sólo puede hacerse si hay espacio para ello y se incluye en el modelo al menos una incógnita real. En un modelo puramente deductivo, no hay lugar para la novedad, la sorpresa o la perspicacia, lo que no significa que el proceso deductivo quede obsoleto. A través de la deducción podemos descubrir lo que no funciona, y esto a su vez es una forma importante de perspicacia. Consideramos que la relación entre la "abducción-inducción" remodelada y la deducción es dialéctica. Esto significa básicamente la síntesis de ambos procesos: el de generación de conocimientos a partir de información individual y el de aplicación de los conocimientos a sucesos conocidos y nuevos con el objetivo de probar críticamente los conocimientos de forma empírica según todas las reglas del arte. La conclusión abductiva del caso individual al caso individual es asumida por este proceso dialéctico y formulada como un paso parcial esencial hacia reglas más abstractas y generales (idealmente enunciados de derecho). Pero en la actualidad nadie cree seriamente en la posibilidad de que los enunciados reales del derecho en las ciencias sociales puedan alcanzar el nivel de las ciencias naturales, que están simplemente estructuradas en comparación con las ciencias sociales. Por lo tanto, se trata más bien de una orientación, pero desde luego no es la norma que cabe esperar. Por encima de todo esto está la capacidad de los

investigadores para cambiar sus creencias cuando hay buenas razones y no en caso contrario. Pólya (1954a, p.8) resume este proceso,

> „First, we should be ready to revise any one of our beliefs.
> Second, we should change a belief when there is a compelling reason to change it.
> Third, we should not change a belief wantonly, without some good reason."

Aunque la deducción puede utilizarse para explicar de forma transparente y verificable cómo probar y según qué criterios, estos criterios confirmatorios pierden su poder en el contexto de la generación de nuevos conocimientos. Por supuesto, podemos inventarnos algo a posteriori sobre cómo hemos llegado exactamente a nuestras conclusiones. Sin embargo, sería mucho más honesto dejarlo como a menudo tiene lugar, es decir, de sopetón y sólo explicable condicionalmente, pero en el nivel sustantivo, es decir, en la teoría del objeto, formulable de forma muy clara e inequívoca. Hay que entender que la ciencia necesita necesariamente estos componentes inciertos, pero su presencia no la hace menos seria y comprensible. Lo que se está produciendo aquí es la transición de la artesanía al arte, y esto se caracteriza por la creatividad y la intuición y sólo puede comunicarse y explicarse hasta cierto punto. La artesanía, en cambio, puede explicarse y enseñarse, pero el arte ya no. Pero el arte se puede experimentar y no requiere ser un genio, sino que es potencialmente accesible a todo el mundo. Sin embargo, esto no significa que todo el mundo vaya a convertirse en un genio de la ciencia, simplemente se requiere un nivel muy alto y el dominio del oficio es lo primero. Esta red de relaciones de los términos tratados debe descubrirse en uno mismo y en el examen de la propia materia. Sin embargo, dado que los descubrimientos deben examinarse críticamente de forma probada gracias al proceso deductivo, la ciencia no pierde por ello nada de su seriedad. Al contrario: sólo puede ganar.

Tarea 2.1: Revisión crítica

La tarea de los lectores en este punto es pensar críticamente sobre esta propuesta y revisarla para ver si es apropiada.

Sin embargo, la ciencia (teoría) es algo más que sacar conclusiones formales. Y la cognición no sólo va a veces por otros derroteros. Con la introducción de la creatividad y la intuición hemos dado los primeros pasos; la próxima sección aborda otros aspectos.

2.5 Excursus – Jugar, hacer y fabricar

Como inspiración, añadimos una cuarta a las tres posibilidades formales de inferencia: *jugar, hacer y fabricar*. Para aclarar las razones, pasamos a algunos ejemplos y consideraciones relacionadas en forma de digresión. Con ello se pretende más inspirar que aportar una "prueba" lógico-formal de la validez de nuestros pensamientos.

Después de todo, sería demasiado fácil creer que la ciencia procede realmente en la realidad del modo sugerido por una conclusión formal en abstracto. Eso sería tan ingenuo como creer que un investigador conductista se movería fuera del ámbito de las leyes del aprendizaje específicamente postuladas. Basta con sentarse un rato, cerrar la boca, cerrar los ojos y escuchar o mirar dentro de uno mismo. Se pueden observar tantos procesos de pensamiento diferentes, deseos, expectativas, opiniones, codicia, odio, negatividad, aburrimiento, pero también compromiso, empatía o amor. Sin embargo, es muy probable que ni siquiera apreciemos esta observación interna, porque enseguida nos quedamos atrapados en cualquiera de estos procesos de pensamiento y olvidamos de inmediato nuestra tarea real de observar. Todos estos fenómenos internos se alternan continuamente en un flujo que cambia constantemente, es decir, todo el día, toda la

noche, día tras día, en el trabajo, de vacaciones, cepillándonos los dientes, ... hasta que muramos. Aunque "sólo" estemos haciendo ciencia, en ese momento este proceso monológico infinito interior no se detiene ni cambia. Esto es independiente del hecho de que en la superficie de nuestra conciencia no solemos ser conscientes de estos procesos más profundos.

Ahora nos preguntamos de nuevo: ¿qué es la ciencia? ¿Qué son los métodos científicos? Nuestros impulsos inconscientes tienen muy poco que ver con las conclusiones formales. A este nivel tan profundo, la ciencia no tiene sentido. Allí ya no existe. Sólo existen sensaciones físicamente palpables ante las que reaccionamos constantemente con rechazo o apego (Goenka, 1991). Pero estas agitaciones interiores tienen un efecto vehemente en nuestro sentir, pensar y actuar a nivel consciente. Y, por tanto, afectan directamente al proceso científico en todas sus fases. Esto se sabe no sólo desde Sigmund Freud (1856-1939), Wilhelm Reich (1897-1957) o la neurociencia moderna.

El conocido ejemplo de August Kekulé (1829-1896), que sentó las bases de la moderna teoría estructural de la química al descubrir la estructura del anillo bencénico, es un ejemplo popular de creatividad en el proceso científico (Hussy, 1986). Sin embargo, la intuición no llegó a través de conclusiones lógicas, sino a través de un sueño y, por tanto, mediada por partes de la propia dimensión en la que se basa nuestra humanidad, pero de la que precisamente no somos conscientes. Esto no fue ni "metódicamente controlado" ni el resultado de una conclusión ni nada de lo que comúnmente llamamos científico. Y, sin embargo, se produjo y tuvo consecuencias espectaculares. Sin embargo, esto fue posible gracias a los numerosos trabajos teóricos preparatorios de Kekulé: sus ideas no surgieron de la nada. Pero la integración se produjo a un nivel inconsciente.

La ciencia es, por una parte, un intento de comprender la vida según ciertas reglas fijas y, si es posible, infinitamente eficaces, y de apropiarse del mundo a través de ellas. Intentamos comprender, explicar, predecir e influir. Por otra parte, apenas sabemos dónde o en qué profundidad dentro de nosotros "surge realmente la ciencia".

Las filosofías seculares, las religiones y los sistemas espirituales suelen perseguir objetivos similares a los de la ciencia: comprender el mundo que nos rodea y a nosotros mismos en él. En el caso de la religión y la espiritualidad, suele haber un componente adicional de otro mundo, que omitiremos para nuestra digresión aquí. Así que en lugar de analizar las diferencias entre religión, espiritualidad o misticismo y ciencia, podemos examinar las similitudes. No sin segundas intenciones, Feyerabend (1976) señala que si bien hemos logrado la secularización – es decir, la separación de la Iglesia y el Estado – en el mundo occidental, no hemos logrado la separación de la política y la ciencia. La ciencia adquiere a menudo rasgos religiosos al ser "creída", ya sea por las palabras de otros científicos, su reputación, el intelecto propio o ajeno, las máquinas, los instrumentos, los grandes conjuntos de datos, las ideas interesantes y plausibles, los análisis y, hoy en día, la informática, la IA o cualquier tweets, los bots y las apps. *Ningún científico experimenta directamente ninguno de los conocimientos científicos de que disponemos*. Todos los resultados así obtenidos se generaron y recibieron de forma mediada, incluso por los científicos que realizaban los estudios. La creencia en la ciencia se basa en la capacidad de generar modelos del mundo que son falsos per se para conectarlos con otra creencia: en los datos. Aquí la conexión con el amplio campo del constructivismo en todas sus manifestaciones es evidente. Y la propia creencia se transporta a través de esas conclusiones que proporcionan el marco básico para estructurar los procesos de pensamiento posibles y permisibles dentro de la ciencia. Así pues, es precisamente anticientífico experimentar las cosas directamente en uno mismo y creer que son ciertas o declararlas científicas. Si, por el contrario, independientemente de nuestra estructura somato-mental, se realiza un estudio sobre el mismo tema con el mismo resultado, los resultados se consideran verdaderos en el sentido de demostrados según Popper o según hacia quién o hacia qué nos orientemos en términos de teoría científica. Esto nos impide investigar directamente y sin prejuicios el proceso que está teniendo lugar en realidad. Todo es un modelo o una construcción. Al fin y al cabo, no significa que no nos hagamos daño al chocar contra una pared sólo porque la física cuántica afirme concluyentemente que en realidad "no hay nada ahí". Lo que ocurre en caso contrario se describe con humor en la película "Men Who Stare at Goats" (2009).

El poder de las creencias y los constructos que las acompañan fue especialmente impresionante de observar en la entrada de la neurociencia en el discurso científico. Sólo se necesitaba una imagen fMRT coloreada del cerebro (los datos brutos no están coloreados, sino los colores sólo se crean mediante

subdivisiones realmente arbitrarias) a través de un proyector en la pared y - voilà - todos los presentes en la sala quedaron impresionados por los informes y resultados que se avecinaban. En los inicios de la neurociencia de la imagen, flotaba en el aire la expectativa de que seríamos capaces de "descodificar" a los humanos y comprenderlos por fin, igual que los físicos intentan hacer con el universo (palabra clave: "teoría de la unificación") o los genetistas con el ADN. La realidad parece entonces mucho más modesta, y en el caso de la neurociencia puede incluso que debido a una metodología estadística defectuosa (Bennettt, Baird, Miller, y Wolford, 2009; Kriegeskorte, Bodurka, y Bandettini, 2008; Eklund, Andersson, Josephson, Johannesson, y Knutsson, 2012; Wow, Krishnan, y Wager, 2014; Eklund, Nichols, y Knutsson, 2016a) gran parte de los resultados de las últimas décadas son inválidos o al menos muy inciertos. También sabemos poco sobre la relación entre las regiones cerebrales activadas y la experiencia real, los llamados qualia, es decir, cómo se forma neuronalmente la experiencia subjetiva. Constituyen un puente entre la investigación del cerebro y la filosofía, y no está nada claro si son tan personales y privadas como se supone. Por otro lado, la investigación del cerebro aún no ha llegado muy lejos en la descodificación de la conciencia.

Porque incluso si una determinada región del cerebro parece estar "más activa" durante ciertas actividades, esto sigue sin permitir llegar a la conclusión causal de que algo o todo lo relevante ocurre exactamente ahí y en ningún otro lugar. Además, la propia metodología de la fRMT limita lo que se puede estudiar y cómo. Actualmente, es imposible estudiar las actividades reales y naturales de la vida en movimiento y la correlación de la actividad cerebral y la experiencia es, en el mejor de los casos, correlativa, pero nunca lógicamente causal. En la medida en que la metodología de investigación determina los objetos que pueden estudiarse, esta influencia masiva o sesgo del objeto así introducido debe tenerse en cuenta en todas las conclusiones. En la práctica, esto no se hace: un vistazo a cualquier artículo de neurociencia demuestra que la metodología de investigación subyacente no se incluye como correctivo en las conclusiones.

Otro estudio de caso apunta a los resultados de la investigación en el campo de la meditación y sus beneficios para el ser humano. Esta línea de investigación ha sido impulsada principalmente por los enfoques de la terapia cognitivo-conductual. Existen beneficios positivos a nivel emocional, en la gestión del estrés, con respecto a parámetros fisiológicos, para uso terapéutico, etc. Pero el núcleo real de la meditación, el desarrollo de la comprensión de la naturaleza del cuerpo y la mente y la subsiguiente transformación individual (diferente), no se aborda o, si se hace, se aliena fuertemente. ¿Por qué? Con sentido común podemos identificar dos razones: En primer lugar, la investigación de la propia mente requiere mucho esfuerzo, según la sabiduría ancestral, un cierto estilo de vida, mucha disciplina y la necesidad de desprenderse de herramientas convencionales ajenas a nosotros mismos. ¿Qué científico está dispuesto a pasar varias décadas de entrenamiento mental intensivo antes de llegar a un nivel que le permita mirar a los ojos a su objeto de estudio? Además, no se puede forzar el progreso en el campo de la meditación. Todo tiene un impulso natural propio. Como otra razón, todas las tradiciones de meditación dicen que la mente sólo puede examinarse con la propia mente, no con un instrumento externo, una máquina o incluso un cuestionario. Además, la meditación es una inversión en el futuro a largo plazo. Esto requeriría un estudio a muy largo plazo para poder trazar realmente cambios más profundos. Sin embargo, con un instrumento externo sólo pueden examinarse los efectos, no el proceso en sí. Y la práctica de la meditación, además, cambia al practicante. Todo el proceso "científico" se desplaza completamente hacia el interior, ya que es el único lugar donde podemos percibir algo directamente y sin intermediarios. Es una "subjetividad objetiva" que cambia de un momento a otro. En general, todo esto se considera menos científico. Pero, ¿eso lo hace menos real e innovador, o incluso menos transformador? ¿Tenemos siquiera medios y criterios adecuados para responder a esta pregunta? Recordemos la opinión budista de que la cognición "verdadera", es decir, la que cambia la vida, tiene lugar exclusivamente en el sujeto y no fuera de él. Sin embargo, se trata de un acceso directo a uno mismo y dentro de uno mismo, donde no hay separación de percepción, análisis e interpretación. Todo esto coincide: la percepción es cognición y la cognición desemboca en la percepción. Y si esto no tiene lugar, no es (todavía) cognición.

Volvamos ahora a las afirmaciones hechas al principio sobre la eficacia de los procesos inconscientes en el proceso científico, que no pueden investigarse científicamente de la forma convencional, a lo sumo sus imágenes (por ejemplo, declaraciones verbales, reacciones no verbales, parámetros fisiológicos, acciones, etc.). Aquí, la meditación sería el instrumento por excelencia para investigar exactamente esto: el mundo espiritual interior y sus cambios. Pero, desde luego, esto no puede realizarse a lo largo de un proceso

científico tradicional. Necesitaríamos criterios totalmente nuevos para garantizar de algún modo la validez de los resultados. Y también el instrumento de recogida ha cambiado obviamente y sobre la documentación de los resultados ni siquiera sabemos si es posible y cómo. Desde esta perspectiva, la ciencia tal y como se practica actualmente parece incompleta. Pero en eso consiste precisamente la ciencia – para adquirir conocimientos. Obviamente, esto también es posible con total independencia de los procedimientos científicos actuales. No olvidemos que muchas de estas tradiciones y técnicas son más antiguas que los orígenes más antiguos de nuestra ciencia actual.

¿Las aproximaciones al mundo descritas ahora serían científicas según los criterios de la ciencia actual? Probablemente no, porque ni siquiera está claro si este conocimiento puede comunicarse adecuadamente y de qué forma. Por eso no es de extrañar que las neurociencias, por ejemplo, pretendan investigar su objeto de estudio partiendo del supuesto de que los procesos mentales pueden reducirse por completo a actividades cerebrales materiales. No hay pruebas de ello y probablemente sea incluso un problema falsable, ya que se necesitaría una metodología reconocida como equivalente para investigar la conciencia independientemente de la neurociencia y compararla críticamente con los hallazgos de ésta. La única opción que queda es investigar la mente con la mente, y eso no suele considerarse muy científico. Metodológicamente no hay otra forma de hacerlo, porque de lo contrario se introducen incógnitas para las que no se dispone de instrumentos de investigación. Y estas incógnitas podrían afectar masivamente a las ecuaciones estadísticas, hasta tal punto que los resultados podrían estar plagados de la máxima incertidumbre. Por tanto, no hay pruebas definitivas de que las actividades cerebrales sean exclusivamente de naturaleza material, lo que, por implicación, no significa negar el papel biológico del cerebro. Sin el cerebro, nada funciona, pero sin el corazón, tampoco funciona nada. De ello se deriva a menudo una falacia lógica, a saber, que las actividades humanas pueden o incluso deben reducirse de repente a puras actividades cerebrales. Un ejemplo es el término "el cerebro adicto" en la investigación de la adicción, un término popular en el título de artículos de revistas relevantes (por ejemplo, Robinson y Berridge, 1995). Sin embargo, no es el cerebro el que es adicto, sino toda la persona. El resto, es decir, el cuerpo completo, queda fuera de la ecuación.

La cuestión de si se ve fuera del cerebro, por ejemplo en las células, y qué papel desempeña la memoria celular, también sigue sin respuesta. Esto cambiaría permanentemente la visión de la adicción y el papel del cerebro, por ejemplo. También tendría un enorme impacto en la industria de los trasplantes, porque entonces no sería simplemente un órgano lo que se trasplanta, sino un órgano y elementos asociados de la persona de la que procede el órgano. No sabemos si es así, pero tampoco podemos descartarlo. Reconocido y concreto objeto de investigación es ahora el sistema nervioso abdominal, de modo que la atención ya no se centra enteramente en el cerebro.

La confusión entre cerebro y ser humano completo fue criticada hace tiempo por Bennett y Hacker (2003, p.3) en el prefacio de su libro sobre los fundamentos filosóficos de la neurociencia:

> „The ascription of psychological — in particular, cognitive and cogiative — attributes to the brain is, we show, also a source of much further confusion. [...] It is the animal that preceives, not parts of the brain, and it is human beings who think and reason, not their brains."

A la inversa, veamos la "neuroinvestigación" concreta de personas tan conocidas como el investigador del aprendizaje y el cerebro Manfred Spitzer, del ZNL Ulm (Alemania). ¿Trabajan aquí exclusivamente con métodos de imagen? No necesariamente, porque los métodos y tipos de datos utilizados abarcan un amplio campo que parece incluir, además de los constructos biológicos y neuropsicológicos, la sólida investigación básica de la psicología cognitiva que se practica desde hace décadas. Esto pone de relieve que los conocimientos reales en el campo de la neurociencia no se obtienen precisamente sólo con métodos de imagen, sino que requieren un amplio espectro de fuentes de datos. Así, a pesar de la riqueza de los hallazgos en neurociencia, la fe más bien ciega en las técnicas de imagen parece ser mayor que su utilidad real o la obtención de conocimientos más profundos. Es evidente que la dimensión de la fe está presente sin que necesariamente se reflexione sobre ella.

Pero, ¿son ya los intentos descritos desde el campo de la lógica y el pensamiento todo lo que podemos hacer para generar conocimiento? Dejemos a un lado las excursiones a los reinos de la espiritualidad, la meditación y un plano espiritual potencialmente puro. Como ya se ha dicho, aquí carecemos de posibilidades

de evaluación seria y no queremos especular demasiado. Por otra parte, ahora sale a la luz una visión mucho más profunda del filósofo David Humes. No sólo se ocupa de la cuestión "superficial" de la inducción como conclusión errónea, por la que es comúnmente conocido. Le preocupa el hecho de que la totalidad de la experiencia humana en sí misma no nos permita deducir de ella leyes siempre eficaces que no cambien en el futuro. De este modo, Hume hace hincapié en la mutabilidad del mundo como característica central del mismo. Nuestra experiencia pasada en el día de hoy no nos permite deducir infinitas afirmaciones sobre el futuro venidero. Ahora bien, en cuanto a la diferencia entre formalismo y vida, el filósofo chino Láozí (supuestamente del siglo VI a.C.) nos lo señala justo al comienzo de su escrito principal Dàodéjíng (en traducción de Richard Wilhelm, verso 1, Láozí, 2007, énfasis en la traducción, nota de los autores):

> "El SENTIDO que se puede pronunciar no es el SENTIDO eterno. El nombre que se puede nombrar no es el nombre eterno".

En cuanto sacamos algo del flujo de la vida e introducimos un signo de puntuación, esta imagen ya no es la vida misma, sino precisamente una imagen u otra cosa. No otra cosa expresan las formas budistas de cognición (véase el capítulo 1.3) o la alegoría socrática de la caverna. Esta función cartográfica no es cuestionada por la ciencia. Pero si la ciencia puede tratar la vida exclusivamente en este nivel constructivista, no hay necesidad de explicación o interpretación científica para entender que hay aproximaciones a la vida igualmente válidas, pero que no necesitan explicación científica ni tienen por qué seguir principios y reglas científicas. Si no existe una regla absoluta, la ciencia es un intento justificado y legítimo de describir y explicar el mundo, pero no es el único intento legítimo al que pueden aplicarse estos criterios. Esta cuestión es bien conocida y es ampliada por Feyerabend (1976) y de forma diferente por Kuhn (1973), entre otros. No profundizamos en este punto. Nuestro punto es mirar más allá del hecho de que la ciencia sólo permite una forma limitada de cognición y que ésta no es demostrablemente la forma más elevada que a nosotros como seres humanos se nos permite mirar.

¿Por qué entonces la referencia a jugar, hacer y fabricar como formas legítimas de cognición? En nuestra opinión, este ámbito es precisamente una forma legítima de adquirir conocimientos sobre uno mismo y el mundo al margen de la ciencia y no según criterios científicos. Observemos cómo los niños aprenden y comprenden el mundo antes de ir a la escuela y "aprender" conocimientos formalizados sobre el mundo. Entonces nos damos cuenta de que los niños hacen y aprenden cosas que difícilmente pueden describirse con inducción, deducción y abducción tre fin. Esto es especialmente perceptible a una edad en la que el lenguaje aún se está desarrollando y en la que intervienen las percepciones sensoriales y los objetos (por ejemplo, las reacciones circulares de Piaget). Utilizar aquí términos científicos sólo significa que estamos interpretando algo apropiado para niños desde el punto de vista de un adulto. Pero en realidad deberíamos interpretar desde el punto de vista del niño, cosa que no podemos hacer: Los niños hacen algo y empiezan a reconocerlo. La interpretación adulta es incompleta y, si lo es, sólo es útil para los adultos. No es una explicación válida describir los "procesos reales" desde la perspectiva interior. No sabemos cuáles son los procesos reales. En cambio, el significado puede reconstruirse. Sin embargo, lo que reconstruimos, lo reconstruimos con gafas de adulto para adultos y, por tanto, como desde otra dimensión. Todo esto es una verdad parcial legítima, pero sólo una verdad parcial, no la cosa en sí.

Pasemos de los humanos a los animales. Los animales pueden pasar mucho tiempo con cosas sencillas, como un cubo, un palo u otro instrumento. Juegan" con ella, a propósito (si hay comida de por medio), pero no necesaria-mente. Y aunque no podemos tener ni idea de lo que ocurre realmente en un animal, es obvio que este jugar, hacer y fabricar cambia a los animales y les aporta una nueva forma de cognición, en el sentido de que utilizan el cubo, el palo, etc., lo usan, quizás lo destruyen. No se produce una cognición formalizada o formalizable, como sugieren los tres primeros tipos de inferencia, como se menciona en el ejemplo de los niños. Es probable que esta forma de abstracción cognitiva no sea propia de los animales. Pero, ¿por qué la cognición siempre tiene que implicar abstracción? ¿Realmente necesitamos la abstracción para poder seguir desarrollándonos? Lo que estamos haciendo aquí es generar términos para "procesos" que surgen y desaparecen sin lenguaje ni formalismo. La evolución no necesitó al hombre para hacerlo surgir. Y gracias a Charles Darwin (1809-1882), Jean-Baptiste de Lamarck (1744-1829), Alfred Russel Wallace (1823-1913) y otros teóricos de la evolución, sabemos que en la evolución ninguna persona la controla, ni

hay un objetivo claro fijado por nadie. Esto no excluye la posibilidad de que la evolución sea mucho más interactiva con las exigencias ambientales y menos aleatoria, incluida la mutación genética. El ADN como único factor explicativo no es suficiente. Pero esto requiere más investigación empírica. Por ejemplo, la epigenética, gracias a la creciente investigación, puede explicar cada vez mejor la transmisión de rasgos adquiridos ontogenéticamente, postulada originalmente por Lamarck.

Por último, nos gustaría señalar y cuestionar una forma científica de adquirir conocimientos. Tomemos el proceso de generación de estructuras y quizá incluso de tipos a partir de datos cualitativos. Nosotros mismos ofrecemos orientación y consejos útiles en el capítulo 9 sobre cómo enmarcar este proceso de la forma más metodológicamente controlada posible. Pero veamos más de cerca el proceso, por ejemplo, basándonos en nuestras propias experiencias con el análisis de datos que aquí son específicamente irrelevantes. ¿Funcionan realmente estos procesos siguiendo estricta-mente una secuencia de inferencias inductivas, deductivas y abductivas? Ciertamente encuentremos buenos argumentos (post-hoc) de que esto fue así y en una publicación seguramente podremos justificarlo bien. Por buenas razones, nadie lo cuestionará. Existe un consenso tácito al respecto. Sin embargo, si miramos un poco más dentro de nosotros, surgen dudas: ¿Tenemos realmente el control sobre cómo procede nuestro pensamiento de forma metódicamente controlada? ¿Cómo se nos ocurre algo nuevo, que obviamente es necesario para la cognición (por ejemplo, la formación de tipos), a menos que estemos reproduciendo algo que ya hemos visto o leído en otra parte? Porque entonces esto no sería más que un trabajo creativo de memoria. ¿Pero tal vez lo sea? ¿No será que de repente "reconocemos" más bien como Kekulé en un sueño o como el "relámpago" de Peirce, pero sin saber realmente cómo se ha producido exactamente en nosotros? A pesar de todos los supuestos esfuerzos científicos, ¿podemos estar seguros de que esto ha sido consecuencia de "conclusiones estrictamente científicas"? Nuestra propia experiencia -sigamos con el ejemplo de la mecanografía (Gürtler, 2005)- se parecía a esto: Tras meses de trabajo de codificación (véase el cap. 9.1) y comparación sistemática de casos (véase el cap. 9.1 o 9.5.6) utilizando el análisis implicante (véase el cap. 12; Huber & Gürtler, 2012) y con el objetivo de llevar los casos a una estructura tipológica, no surgió ningún hallazgo controlado en el proceso científico consciente. Y sin embargo, un día bastó con echar un vistazo a los datos y la tipología apareció "como por sí sola y sólo había que ponerla por escrito". Este último acto necesario para la cognición no era controlable, ni en el tiempo, ni en el contenido, ni en la estructura. Y se parecía mucho a lo que informó Kekulé.

La ciencia gira en torno al respectivo control metodológico propagado y es precisamente esto lo que nos gustaría cuestionar. El control como concepto nos parece aquí ilusorio. Queremos dejar deliberadamente las preguntas que surgen en este punto o en para todos los lectores como inspiración y estímulo. De todos modos, la respuesta no es fija y no tenemos necesidad de "demostrar" que es así. ¿Cómo podríamos hacerlo? ¿Con métodos científicamente aceptados? Entonces nos estamos contradiciendo. Para nosotros, el problema es de naturaleza dialéctica: una parte de nuestra mente trabaja ciertamente en función de lo que hemos aprendido sobre una base científica. Sin embargo, no suponemos que ésta sea la parte más grande en nosotros y sabemos igual de poco cómo funciona "correctamente" esta parte en realidad. Por el contrario, parece ser una parte menos significativa en nosotros, que probablemente no tiene ni idea de lo insignificante que es en realidad. Y esto da lugar a muchos (auto)malentendidos. Si entonces llegamos a algún tipo de realización de alguna otra manera – y aquí nos saltaremos los detalles –, esta pequeña parte se apropia de los resultados e, ignorante de las verdaderas circunstancias – que, de todos modos, probablemente estén cambiando constantemente –, los hace pasar por sus propios logros.

En resumen, tenemos un problema de ego en la ciencia. Quizá podamos superarlo explorando todas las áreas de nuestro interior y aprendiendo a integrarlas entre sí. Suena sencillo, pero es difícil de aplicar. Es aún más importante estar abierto a nuevas formas de conocimiento. Y también es importante entender bien los hallazgos conocidos de la ciencia sobre la ciencia. Por lo tanto, pasamos ahora a lo que consideramos las ideas más importantes de las teorías actuales de la ciencia.

Capítulo 3
Hitos de la Filosofía de la Ciencia

»When you develop your opinions on the basis of weak evidence, you will have difficulty interpreting subsequent information that contradicts these opinions, even if this new information is obviously more accurate.«

The Black Swan: The Impact of the Highly Improbable.
Nassim Nicholas Taleb, 2007

En 2007, el matemático financiero y estadístico Nassim Nicholas Taleb (1960-) escribió un libro sobre sucesos impredecibles (Taleb, 2007), que pueden ocurrir a pesar de una probabilidad mínima y tener así un impacto duradero en el curso de la historia. El autor acuñó el término *antifragilidad* para describir precisamente esta imprevisibilidad y amplió su alcance a muchos ámbitos cotidianos de la vida – como la salud, la educación, la política, la economía y la cultura, por citar sólo algunos. Al mismo tiempo, enfureció a la comunidad estadística al afirmar que en tiempos de crisis y ante el riesgo (financiero), los estadísticos tienden a esconderse tras complicados modelos mate-máticos. Independientemente de lo acertadas que sean las críticas de Taleb en casos concretos, su obra estimula el pensamiento independiente. Esto es especialmente relevante en el contexto de la ciencia, ya que la creencia en la ciencia puede adquirir connotaciones religiosas, por lo que el término *psicología de masas* (Reich, 1933), bastante anticuado, bien podría aplicarse en la argumentación. En consecuencia, observa Taleb (2007, p.192), en el pasado era aparentemente más probable ir juntos en la dirección equivocada que solos en la correcta. Y hemos heredado genes de esa época:

> "Those who have followed the assertive idiot rather than the introspective wise person have passed us some of their genes. This is apparent from a social pathology: psychopaths rally followers."

¿Por qué hacemos esta introducción ahora, cuando hablamos de hitos en la filosofía de la ciencia? Pues bien, precisamente cada uno de estos hitos esboza los esfuerzos de personas individuales innovadoras y valientes que añadieron una nueva pieza al rompecabezas de la búsqueda de la verdad de los hechos. Lo que todos ellos tienen en común, en mayor o menor medida, es que sus puntos de vista fueron en parte rechazados inicialmente, en parte aceptados pero no necesariamente adoptados o, como en el caso de Paul Feyerabend, provocaron las críticas más masivas, incluso hostilidad personal. Y ello únicamente porque los respectivos representantes publicaron sus opiniones en forma de texto y, además, ¡pudieron justificarlas! Si los respectivos representantes se hubieran guiado por lo que al público en general o a la comunidad científica le gusta oír y espera y, por tanto, le resulta emocionalmente cómodo, probablemente algunos habrían dotado a sus trabajos de un contenido diferente. La ciencia no es un espacio aislado y protegido, sino que se sitúa en medio de diversas responsabilidades políticas y sociales. La responsabilidad es una parte obligatoria de la ciencia, de lo contrario falta el último aspecto de la ciencia – intervenir en la realidad por el bien de todos. La actitud y el respectivo enfoque de la responsabilidad sociopolítica se reflejan en la acción.

Dentro de la filosofía de la ciencia hay muchos representantes, de los cuales vamos a esbozar a continuación los más importantes para nosotros con sus principales posiciones. Para nosotros, la pregunta que nos guía es "errar" y sacar conclusiones, así como "¿hay algo nuevo?". Esto no tiene nada que ver con si trabajamos con datos cuantitativos o cualitativos, que es el siguiente tema de este libro. Las posiciones son bastante inespecificas en cuanto al tipo de datos y el diseño, pero esenciales en sus afirmaciones básicas e igualmente pertinentes para los diseños cuantitativos y cualitativos.

Para nosotros, la mejor panorámica en lengua alemana de las teorías de la ciencia se encuentra en la obra en varios volúmenes "Problemas y resultados de la filosofía de la ciencia y de la filosofía analítica" de Wolfgang Stegmüller (1923-1991) y en otras publicaciones de Stegmüller (por ejemplo, 1975 sobre el problema de la inducción), que en nuestra opinión es y sigue siendo inigualable (¿inalcanzable?) en cuanto a precisión, lenguaje, amplitud y comprensibilidad en el ámbito de la lengua alemana. Siempre suscita entusiasmo cuando se lee.

Los filósofos de la ciencia especialmente influyentes para nosotros son David Hume (1711-1776), Karl Popper (1902-1994), Imre Lakatos (1922-1974), Thomas Samuel Kuhn (1922-1996) y Paul Feyerabend (1924-1994). Todos ellos han posibilitado una perspectiva diferente del cientificismo y su contexto, de modo que juntos producen una imagen presumiblemente realista de cómo funciona y debería funcionar la ciencia y por qué es así. Las siguientes piedras angulares de sus ideas y argumentos son, por supuesto, sólo una pequeña muestra del fondo de a?muchos filósofos y epistemólogos - adaptado a los requisitos de este libro.

3.1 Hume

De Hume podemos aprender que no se pueden extraer conclusiones que amplíen el contenido mediante el conocimiento inductivo (problema de la inducción, Stegmüller, 1975). Esto significa que el pasado y toda la información, obser-vaciones, etc. recogidas hasta ahora sobre la base de nuestras experiencias sensoriales no dan ninguna garantía de lo que nos deparará el futuro. El futuro puede ser diferente: las cosas cambian. Se puede decir así: Sólo podemos evaluar la información que poseemos, y la del futuro, mientras el futuro siga siendo futuro, no nos es accesible. Pero aún más importante, según Hume, nuestros sentidos son la única fuente principal de conocimiento y en principio no tenemos ni idea de lo que ocurre realmente en el mundo exterior. Se trata de una visión bastante asombrosa de la cerrazón de los sistemas biológicos para el siglo XVIII (Maturana y Varela, 1984) y, de hecho, es una continuación perfecta de Sócrates y Platón. Más tarde, el constructivismo se basa en el mismo punto de vista y los enfoques sistémicos también se basan todos en la suposición de que las personas se perciben principalmente a sí mismas – y el entorno sólo mediado a través de estímulos sensoriales y órganos de los sentidos, pero no directamente. Por el contrario, Hume se oponía a las conclusiones inductivos (Stegmüller, 1975, Gelman y Shalizi, 2013).

3.2 Popper

Popper (1943) es particularmente importante cuando se trata de replicar estudios para obtener conocimientos sostenibles. Sólo unos pocos lo hacen y no les gusta publicarlo, ya que estas réplicas se asocian con el aburrimiento y la falta de novedad, lo que se considera estereotipadamente como falta de cientificidad en la comunidad científica. Tenemos que agradecer a Popper la enfática exigencia de no (sólo) buscar "pruebas" positivas que respalden las propias suposiciones, sino de investigar activamente la contra-información que pueda echar por tierra la propia construcción de hipótesis. Esto debe aplicarse rigurosamente, se denomina *falsificación* y conduce directamente a la exigencia de replicar los resultados de la investigación. Popper siguió las ideas de Hume sobre el problema de la inducción, las radicalizó y dudó de la posibilidad de la inducción per se (véase el capítulo 12). Para él, la generalización de casos individuales al nivel de una teoría general no era más que una ilusión. En cambio, todas las conclusiones científicas son, en última instancia, deductivas. En consecuencia, criticó el positivismo lógico o empirismo (neopositivismo). En el racionalismo crítico de Popper, la deducción ocupa, pues, un lugar central. Aquí, las teorías pueden y deben inventarse de forma creativa. Posteriormente, se someten a pruebas empíricas críticas junto con las predicciones (teoremas básicos) que pueden derivarse de ellas. El resto sigue un proceso de selección evolutiva, por así decirlo, de modo que la teoría que ha sido repetidamente probada empíricamente de forma crítica, pero no falsada en el proceso, se considera *mejor probada*, pero no demostrada. Para Popper, este proceso es una aproximación a una verdad, pero sin poder determinar la propia aproximación – en sentido absoluto. A la

inversa, de ello se deriva la exigencia de que las teorías estén libres de contradicciones, ya que de otro modo no serían falsables y el progreso se haría así imposible según Popper. Porque sin comprobación empírica crítica, no hay falsificación. Hay mucho más que aprender de Popper, pero su incansable voluntad de examinar críticamente y modificar su propia obra es ejemplar. Lo que también podemos aprender del trabajo de Popper es que siempre podemos equivocarnos, por ejemplo, al intentar falsar, ya que la falsación no es, en última instancia, superior a la inducción o la abducción, sino que utiliza los mismos métodos científicos. La verdad es relativa por naturaleza.

3.3 Lakatos

El punto de partida de Lakatos (1978) es que todos los datos están cargados de teoría. No hay observaciones puras aparte de las ideas teóricas. Y las ideas teóricas compiten entre sí, casi en un proceso evolutivo en el que la que "sobrevive" demuestra su valía. La teoría probada tiene un exceso de contenido, lo que no significa que aquí se produzca una aproximación real a la verdad. Lakatos nos señala que – al intentar aplicar Popper (es decir, buscar contrapruebas o falsificaciones) – también podemos equivocarnos. Tanto si buscamos pruebas positivas como negativas, el error potencial siempre está incluido. En consecuencia, Lakatos ha creado un recipiente con los *programas científicos* dentro del cual las nuevas teorías se protegen primero de una falsificación demasiado rápida en el núcleo ("core"), porque, por ejemplo, el propio proceso de falsificación fue defectuoso o la teoría tiene inexactitudes pero es fundamentalmente válida, pero sólo se probaron las inexactitudes. Del mismo modo, podría ocurrir que la teoría investigada fuera "correcta", pero el intento de comprobación que se llevó a cabo fuera en sí mismo "incorrecto". En consecuencia, el trabajo empírico queda relegado a la periferia para poder evaluar seriamente a largo plazo y sobre la base de muchos estudios si una nueva teoría se demuestra empíricamente o no. Las teorías complejas difícilmente pueden falsarse con un solo experimento, del mismo modo que un solo estudio no puede aportar una prueba completa. En realidad, básicamente las teorías complejas no se pueden falsificar ni validar. Sin embargo, el enfoque núcleo-periferia evita que una teoría nueva y prometedora se descarte por completo al primer experimento fallido o investigación empírica fallida. En psicología, por ejemplo, el programa de investigación "teorías subjetivas" se concibió siguiendo estas líneas (Groeben y Scheele, 1977, Groeben, 1986).

3.4 Kuhn

Kuhn (1973), por su parte, nos ayuda a comprender mejor que la ciencia no lleva una vida propia descontextualizada, sino que está hecha por el hombre dentro de una comunidad social heterogénea. Por tanto, las nuevas teorías no se adoptan necesariamente porque sean mejores en el sentido de aproximarse a una verdad, sino porque los tiempos cambian. Las teorías antiguas se sustituyen por otras nuevas, sin que éstas se acerquen necesariamente a la verdad. Un nuevo *paradigma (estilo de pensamiento)* permite resolver problemas que el anterior no permitía – sin embargo, según Kuhn, esto no legitima la cuestión del progreso en el sentido de aproximación a la verdad. Aquí es donde Kuhn difiere de Popper. Vista de forma abstracta y durante largos periodos de tiempo, la ciencia se adentra así en el terreno de las "modas", aunque Kuhn no utilice este término. Así, es habitual en muchas disciplinas que determinadas construcciones teóricas (paradigmas) ocupen el centro de la escena durante determinados periodos de tiempo. Luego son sustituidas por otras, sin que esto pueda explicarse racionalmente a partir de las propias teorías. Posiblemente, sin embargo, la totalidad de los problemas científicos, las explicaciones alternativas existentes, las corrientes sociales y la evolución podrían explicar tales cambios. En este sentido, las posiciones de Kuhn no son sólo epistemológicas o filosóficas, sino al mismo tiempo siempre sociológicas. Sin embargo, Kuhn va más allá y analiza la ciencia y sus desarrollos bajo el aspecto del darwinismo. Aquí Kuhn compara la evolución de los organismos con la evolución de las ideas científicas (1973, p.18s.):

> "Y todo el proceso puede haber procedido de la forma en que ahora suponemos que ha procedido la evolución biológica, sin la evolución, sin la ventaja de una meta bien determinada, una verdad científica supra-temporal, fija, de la que cada nueva etapa en el desarrollo del conocimiento científico es un mejor reflejo."

Los términos *crisis* e *inconmensurabilidad* son centrales en la posición de Kuhn (ibid., cap.VII y p.209 . respectivamente). El cambio de paradigmas comienza con la crisis, porque se producen incoherencias dentro de un paradigma que éste no puede explicar. El cambio al nuevo paradigma se produce sobre una base irracional, ya que, por lo general, el nuevo paradigma aún no ha sido suficientemente investigado empíricamente. Por lo tanto, cualquier avance como aumento del conocimiento a través del nuevo paradigma no está justificado. La comparación de los paradigmas en liza también demuestra que utilizan lenguajes cualitativamente distintos y que, por tanto, son incomparables (inconmensurables). Kuhn ha sido acusado desde diversos frentes, entre otras cosas, de vaguedad en su elección de términos. Sin embargo, hay que subrayar que Kuhn se adentra en la complejidad de unos hechos difíciles de explicar exhaustivamente sin caer en una posición completamente relativista. Al mismo tiempo, se aventura a salir de los estrechos límites de las consideraciones filosóficas y examina la relación entre los científicos y la sociedad. Además, aborda con gran detalle qué es realmente la ciencia (ibid., cap.XIII); y del mismo modo examina la mutabilidad como una figura central de la ciencia. Sin Kuhn podríamos tener la idea de examinar la ciencia sólo dentro de la ciencia, y no simplemente dar un paso al lado.

3.5 Feyerabend

Feyerabend (1976) como último representante tuvo el coraje de pensar consecuentemente Popper, Lakatos y Kuhn más allá y cuestionarlo *todo*. Además de la ciencia, existen muchos otros bienes culturales humanos que también pueden utilizarse como fuentes legítimas de conocimiento. Y luego pueden aplicarse otros criterios y uno no es comparable con el otro. No sólo hay que pensar e investigar empíricamente. Aquí, Feyerabend no deja de introducir polémica en la discusión e, igualmente, de adoptar una postura política para representar adecuadamente las cosas que son relevantes para él: la cognición es relativa y encaja en estructuras históricamente sociales. Desgraciadamente, muchos no entendieron realmente su humor sutil y su compromiso con la ciencia, por lo que se sintió indebidamente enemistado a raíz de esta publicación. Nosotros ?ndemos, sin embargo, que Feyerabend se encuentra entre los más valientes de todos los científicos mencionados, ya que los cuestionó completamente sin negarlos. ¿A quién le gusta oír cuando se le acusa de que todo esto no es tan grande como la propia ilusión quisiera que fuera? ¿Quién se atreve a cuestionar su supuesta superioridad dentro de su propia comunidad? Feyerabend tuvo este valor y sus argumentos sobre la cuestión de la cientificidad no deben descartarse sin más (ibíd., p.408s., cursiva en el original). Esto merece una larga cita.

> "Si queremos comprender la naturaleza y dominar nuestro entorno material, debemos utilizar *todas* las ideas, *todos* los métodos, no sólo una pequeña parte de ellos. Pero la afirmación de que fuera de la ciencia no hay conocimiento -extra scientiam nulla salus- no es más que otro cuento de hadas de lo más conveniente. En las culturas primitivas existen clasificaciones más detalladas de animales y plantas que en la zoología y la botánica científicas actuales, existen remedios cuyos efectos asombran a los médicos (mientras que la industria farmacéutica ya intuye aquí nuevas oportunidades de beneficio) [...] En todas las épocas el hombre se enfrentó a su entorno con los sentidos alerta y una inteligencia fértil, en todas las épocas hizo descubrimientos increíbles, en todas las épocas se puede aprender de sus ideas."

Lo sorprendente aquí es que Feyerabend se refiere a la naturaleza y se desmarca de las profesiones, negándoles la pretensión universal de conocimiento. Se trata más bien de la interacción hombre-entorno a través de la cual surgió la cognición. De este modo, el conocimiento específico de la profesión se contextualiza precisamente a través del ejemplo de la disciplina médica, de modo que la profesión (ejemplo de la medicina) representa sólo una de muchas posibilidades. Feyerabend (ibid.) extiende esta figura a la ciencia como tal:

> "La ciencia moderna, en cambio, no es en absoluto tan difícil ni tan perfecta, como la propaganda científica nos quiere hacer creer. Un tema como la medicina, la física a física o la biología parecen difíciles sólo porque se enseñan mal, porque sus presentaciones habituales contienen demasiadas cosas innecesarias y porque empezamos a estudiarlas demasiado tarde. demasiado tarde en la vida. Durante la guerra, cuando los militares americanos necesitaban médicos con poca médicos con poca antelación, de repente era posible acortar la formación médica a medio año. año (pero los libros de texto correspondientes desaparecieron hace tiempo). En la guerra, la ciencia puede simplificarse. En la paz, su prestigio exige que sea complicado). [...] ¡Cuántas veces la ciencia mejora gracias a aportaciones extracientíficas y se orienta en nuevas direcciones!"

Lo significativo de esta cita es que Feyerabend pasa de la ciencia a la vida cotidiana: la guerra, otras culturas, la educación y la economía o la política. Mientras otros siguen atrincherados en la lógica formal, Feyerabend toma el camino de la ciencia a la vida cotidiana y sus exigencias. Pero es precisamente aquí donde la ciencia es necesaria. Sólo con fines académicos ninguna sociedad puede permitirse dedicarse a la ciencia. Además, Feyerabend adopta una postura crítica y también entiende la ciencia como una evaluación y no como una posición neutral que se retira de la vida cotidiana y la deja en manos de otros. Encontramos algo parecido en la polémica sobre el juicio de valor o el positivismo en las ciencias sociales en los años sesenta, que se remonta a los debates en sociología a principios del siglo XX.

3.6 Conclusión – Filosofía de la ciencia

Así que podemos concluir: *Errar es humano, ocurre siempre y es inevitable*. Algo que prácticamente forma parte de nosotros no debe asustarnos ni cuestionar nuestro trabajo. Más bien puede servir de inspiración. Seamos científicos o no, no podemos salirnos del sistema del error potencial. No hay ni verificación completa de las hipótesis y teorías ni falsificación completa de las mismas. Una falsificación completa sería, pues, una verificación completa bajo signos negativos y, por tanto, estructuralmente igual. Hasta la fecha, nadie ha aportado ninguna prueba razonable y general-mente aceptada de ello en ninguna forma. La única "demostración formal-abstracta lógicamente sólida" parece ser hasta la fecha el Teorema de la Improbabilidad de Gödel (1931). En resumen, afirma que dentro de un sistema hay proposiciones y afirmaciones que formalmente no son ni demostrables ni refutables o, en otras palabras, que un sistema matemático no puede demostrarse a sí mismo. Podría decirse que esto también se aplica a los sistemas lógicos no matemáticos y no formales (véase Jaynes, 2003).

El conocimiento y la cognición están contextualizados y como tales son relativos, pero esto no significa que actuar de acuerdo con estas cogniciones carezca de consecuencias reales, sino todo lo contrario. La acción en el mundo siempre conlleva consecuencias y éstas actúan como causas de otras consecuencias contextualizadas dependientes, etc. La inducción, en particular, es propensa al error, salvo en la imaginación de las matemáticas (inducción matemática), que tiene lugar en un terreno de juego asegurado y conlleva pocas consecuencias para la realidad, salvo quizá una cierta inspiración. En sentido figurado: en principio, un vaso puede volar roto desde el suelo hasta la mesa y volver a montarse perfectamente. Sin embargo, la probabilidad de que esto ocurra es tan baja que nosotros y todas las generaciones posteriores, hasta la extinción de la humanidad, desgraciadamente no lo experimentaremos. Esto está estrechamente ligado a la estructura de nuestro espacio-tiempo (Greene, 2008). Y no hay ninguna certeza de que esto vaya a ocurrir realmente, ni mucho menos de que vaya a ser observado por los seres humanos. En la práctica, se aplica la ley de la entropía, la segunda ley de la termodinámica. No obstante, como demuestran las estadísticas bayesianas, la conclusión inductiva puede aportar un alto grado de conocimiento en condiciones reales y ser útil. Prescindir de él sería peligroso.

Recordatorio 3.1: Cambio

En materia de error, la ciencia no es una excepción, sino parte de lo jerárquicamente superior – ¿absoluto? – regla, y es que las cosas cambian constantemente.

Pero la cuestión no es que podamos equivocarnos y que se produzca ese error, porque eso es trivial. Más bien *tenemos que prestar atención a qué hacemos con esta información básica de la mutabilidad del mundo* y si incorporamos precauciones a nuestro trabajo para tener en cuenta esta misma circunstancia y modelizarla adecuadamente. Poder equivocarse en principio no significa que por ello debamos dejar de lado la ciencia o que sólo sea legítimo un determinado camino, sino sólo que tenemos que trabajar con mucho cuidado y comprobar una y otra vez nuestros resultados para comprobar su coherencia, su constancia a lo largo del tiempo y los cambios que conlleva.

Por último, conectamos la teoría de la ciencia con la vida cotidiana: cocinamos fideos y no sabemos mucho sobre los fideos. ¿Cómo sabemos si la pasta está *al dente*?

- Popper establecería un modelo cognitivo, una teoría deductiva sobre la cocción de los fideos y luego probaría críticamente si los fideos están cocidos. La idea de simplemente cocinar fideos basándose en la experiencia de cocinar fideos y dejándose llevar por la intuición, probablemente no sería para él. Tampoco lo sería la conclusión de generalizar inductivamente esto a todos los tipos de fideos después de haberlos cocinado correctamente una vez por casualidad.
- Hume confiaba en la experiencia y sacaba los fideos del agua hirviendo de vez en cuando para ver si estaban listos. Pero nunca estaría seguro de si esto se aplica a todos los fideos, ahora y en el futuro, tanto si se cocinan como si no.
- Feyerabend tal vez miraría lo que otros cocinan y disfrutaría de esta comida.
- Probablemente, Pólya también probaría los fideos y, si le gustaran, le parecería muy plausible que los demás fideos de la olla también estuvieran cocidos o que simplemente se cocinaran los fideos como él. Y viceversa: si uno está poco hecho, los demás lo estarán con muy poca verosimilitud; y en el futuro aprende de la experiencia a no probar hasta más tarde.
- Los niños primero se queman los dedos con los fideos en el agua caliente y aprenden muy rápido a sacarlos con la cuchara y esperar un poco. Pero luego se divertirán aún más tragándoselos con ruidos extraños y echándoles un montón de ketchup por encima. Y, de todos modos, que los fideos estén cocidos no es cosa suya, sino de sus padres. Sin embargo, si no lo son, podemos estar seguros de que refunfuñarán en voz alta.

Y ese es un buen punto de partida. Creemos que los *métodos mixtos* pueden aportar una contribución importante a este respecto, si se aplican de forma adecuada a cada caso. La herramienta más importante es *el sentido común*.

Parte II

Métodos Cuantitativos

Capítulo 4
La Estadística Clásica

> »Wenn vor Jahren schon die Zahl der Brücken veröffentlicht wurde, die in den nächsten Jahren einstürzen werden, und diese Brücken dennoch einstürzen, ist damit nichts gegen die Statistik gesagt, sondern einiges über die bedauerliche Tatsache, daß die richtigen Zahlen nie von den richtigen Leuten zur rechten Zeit gelesen werden.«
>
> "Si hace años ya se publicó el número de puentes que se derrumbarán en los próximos años, y estos puentes no obstante se derrumban, esto no dice nada en contra de las estadísticas, sino algo sobre el desafortunado
>
> Dieter Hildebrandt, 1927–2013

Las siguientes explicaciones no sustituyen a una introducción a la estadística o a las matemáticas subyacentes. Tampoco tienen el carácter de un libro de texto. Suponemos que nuestros lectores tienen algunos conocimientos de estadística y, en caso contrario, que los adquirirán en forma de estudios o seminarios (por ejemplo, Gelman y Hill, 2007; Eid, Gollwitzer y Schmitt, 2010). No pretendemos que los complejos de problemas descritos sean completos, ya que son mucho más complejos en detalle de lo que podemos discutir aquí en el marco limitado. Algunos temas (por ejemplo, clásico frente a bayesiano, la necesidad de réplicas, la paradoja de Lindley, la significancia, los valores p y la com-probación de hipótesis nulas, ...) han dado lugar a un número inabarcable de artículos científicos de una amplia gama de disciplinas. De ellos, sólo extraeremos los temas centrales que consideramos relevantes para los métodos mixtos.

4.1 Un comienzo

El objetivo general de este capítulo es dar a conocer las distintas formas de estadística y los problemas que surgen en el proceso, para no olvidar el sentido común en la investigación. Colocamos deliberadamente el sentido común jerárquicamente por encima de la estadística, casi como un axioma. En nuestra opinión, se tratan los mecanismos efectivos de las teorías estadísticas y se cuestionan las formas de pensar subyacentes, en la medida en que resulten relevantes para el tema de la "integración de métodos CUAN/ CUAL" o para una comprensión fundamental más allá de la corriente generalmente practicada. En cuanto a la estadística clásica, incluye conceptos de apoyo como el tamaño de la muestra, la comprobación de hipótesis nulas, la signi?cancia, la con?dencia, la fuerza del efecto y la potencia. Los problemas y escollos especiales, así como los conceptos erróneos más comunes y extendidos, se abordan por separado según sea necesario. Prestamos especial atención al papel totalmente sobrevalorado y malinterpretado del concepto de significación estadística basado en el valor p (véase el capítulo 4.3.9.1), a las réplicas de estudios que apenas se encuentran y aún menos se aprecian (véanse los capítulos 4.4.4 y 7.10), o al cálculo del coste total que prácticamente siempre falta (véanse los capítulos 4.3.3.5 y 7.10.2).

Aparentemente, las instituciones de investigación y las universidades pueden permitirse no llevar a cabo una contabilidad de costes completa, aunque su ausencia deje la perspectiva económica completamente fuera de la ecuación. Abogamos por evaluar los beneficios de la ciencia para la sociedad. Se trata de un área difícil, porque la ciencia no sólo debe justificar sus beneficios en términos de un retorno monetario inmediato de la inversión („return of investment“ ROI). Por el contrario, la ciencia debe ser capaz de planificar a largo plazo, lo que significa que a veces los éxitos sólo se hacen patentes al cabo de muchos años. No obstante,

pensamos que esfuerzo y rendimiento deben contraponerse, con el aspecto de fondo de reforzar la ya mencionada cultura del error. El trabajo científico implica que grandes beneficios pueden residir en estudios estadísticamente insignificantes que rara vez se mencionan o publican. Las réplicas son esenciales para poner en perspectiva los éxitos individuales de los estudios e investigar su estabilidad. En sentido estricto, todos estos aspectos pertenecen al contexto de la contabilidad de costes totales. No se trata de exacerbar aún más la ya demencial carrera por los fondos de investigación informando obligatoriamente de cualquier tipo de éxito investigador. Por supuesto, existe el peligro de que esto ocurra, por lo que se necesitan normas del propio mundo académico, y no necesariamente sólo de terceros financiadores, sobre cómo llevar a cabo una contabilidad de costes completa. Un ejemplo sería que un resultado estadísticamente significativo vale tanto como un resultado estadísticamente insignificante si la calidad del procedimiento metodológico y del trabajo de justificación teórica está a un nivel comparable.

> Tarea 4.1: Contabilidad de costes totales
>
> Encuentre un solo artículo científico de cualquier área temática en el que, en el marco de la contabilidad de costes totales, se enumere y justifique que la investigación realizada ha merecido la pena y por qué, y en qué relación se encuentran los costes de inversión con el beneficio. No se trata de investigaciones que hayan dado resultados especialmente buenos. Descubrir que algo no funciona es al menos un resultado igual, ¡si no superior!

En cuanto a las explicaciones sobre la estadística de Bayes (véase el capítulo 6), es cierto que hoy en día no todos los lectores pueden asumir un conocimiento profundo de Bayes y que esta materia no suele enseñarse en las ciencias sociales. Por lo tanto, la atención suele centrarse en la estadística frecuentista y los problemas asociados a ella, como reflejo de las condiciones de la vida real. Sin embargo, llevamos a cabo este debate desde una perspectiva que se aleja de las puras pruebas de significancia hacia el modelfitting y el estudio de los parámetros y su variabilidad (Gelman & Stern, 2006) o siempre aborda la cuestión del método de análisis de datos adecuado (Gigerenzer & Marewski, 2015) y los criterios estadísticos a lo largo de los cuales se pueden tomar decisiones (por ejemplo, si vale la pena aplicar ampliamente una formación pedagógica recientemente desarrollada).

Aunque la aplicación de la estadística de Bayes es matemáticamente muy compleja, siempre se basa en el teorema de Bayes (véase el capítulo 6.4). Esto tiene ventajas e inconvenientes. La desventaja es que puede surgir la impresión de que todos los problemas se abordan siempre de manera bayesiana y que no se reflexiona más sobre si esto puede justificarse en absoluto. En relación con los métodos mixtos, esto significa que hay que pensárselo dos veces antes de decidir qué parece adecuado al caso.

Si queremos saber si los estudiantes tienen más éxito con una nueva política gubernamental en materia de educación, calculamos. Pero si queremos entender por qué funciona en unos sitios y en otros no, ya no calculamos, sino que observamos, hacemos preguntas, vamos al lugar y utilizamos muchos otros métodos de recogida y análisis de datos. Si nos interesa lo que alguien dice realmente y no sólo superficialmente (por ejemplo, en un discurso público), analizaremos el caso cualitativamente. No hay nada que calcular. Sin embargo, si queremos trabajar sobre una cuestión lingüística, las matemáticas pueden volver a ser útiles (por ejemplo, el Procesamiento del Lenguaje Natural (PLN)). Es importante analizar detenidamente el contexto en el que nos encontramos y saber si los métodos mixtos pueden contribuir de forma fructífera a responder de forma exhaustiva a una pregunta de investigación. Lo que nos gustaría desaconsejar – en el marco del creciente debate sobre la Inteligencia Artificial (IA) – es el análisis automatizado de datos a través de la minería de datos incontrolada, la codificación (semi)automática de textos, etc. Gigerenzer y Marewski (2015), por ejemplo, discuten la idealización del análisis de datos universal automatizado a raíz de la estadística bayesiana y el eterno sueño de utilizar un método universal independientemente de las necesidades reales y específicas de cada caso. Los criterios deben seguir seleccionándose manualmente en lo que respecta a los análisis de datos. E incluso si los programas de IA pueden algún día facilitar o incluso sustituir nuestro trabajo, seguimos necesitando entender qué resulta y por qué. No podemos eludir el extenuante trabajo manual de investigación y análisis.

En resumen: las distintas formas de estadística tienen cada una su ámbito de aplicación legítimo. Como se verá más adelante, el enfoque Neyman-Pearson, por ejemplo, se adapta perfectamente a la gestión de la calidad (véase el capítulo 4.3.3.4). En principio, Bayes ya no es necesario. Por otra parte, el enfoque bayesiano es muy ?exible y puede aplicarse a una amplia gama de problemas, por ejemplo, tamaños de muestra muy pequeños (Studer, 1996b, 1998).

Por supuesto, la estadística de Bayes comparte ahora algunos problemas estructurales ya conocidos de la estadística clásica. Entre ellas se encuentra la formulación de criterios y umbrales críticos adaptados a cada caso y referidos al contenido en el caso de los resultados empíricos disponibles. Se trata de cómo deben tomarse las decisiones justificadas en el problema frente a la inferencia. Además, hay consideraciones fundamentales relativas a la selección del propio procedimiento de análisis de datos, que no está nada clara. Tampoco hay consenso sobre si ahora se trata de pruebas o del comportamiento de los parámetros del modelo. Recientemente, también se ha observado que con el renacimiento de la estadística de Bayes está surgiendo un fenómeno comparable a la fijación del valor p de la estadística clásica y el ritual nulo que la acompaña (véase el capítulo 4.3.8): un énfasis excesivo en los factores de Bayes (véase el capítulo 6.8.1) independientemente de la pregunta predominante (Gigerenzer y Marewski, 2015; Gelman y Rubin, 1995; Gelman y Shalizi, 2013). Así, los factores de Bayes pueden derivarse de valores p calibrados y no reflejan un enfoque bayesiano completo, sino que sólo reflejan cambios en las expectativas dados los datos. En concreto, se añade una distribución a priori a los p-valores frecuentistas y los factores de Bayes se derivan mediante calibración (por ejemplo, Sellke, Bayarri y Berger, 2001), y algunos autores incluso piden que este procedimiento sea obligatorio por defecto (Iverson, Lee, Zhand y Wagenmakers, 2009). Si algo así se utiliza ampliamente y sin reflexión, puede llevar en el futuro a que las revistas de repente esperen "enormes factores de Bayes" en lugar de "significancias máximas" y los establezcan como requisito previo para la publicación. Esto expulsaría al diablo con el Belcebú. No sólo siguen implícitamente umbrales críticos, que no se fijan adecuadamente para cada caso, sino de forma generalizada. No se practicó ni se practica otra cosa con los p-valores y uno se pregunta si los científicos pueden aprender de sus décadas de errores. No es el paradigma con el que se trabaja lo que garantiza la calidad de la investigación, sino la flexibilidad con la que se aplica este paradigma en función de la situación.

Si se antepone la generalidad a la adecuación al caso, los resultados seguirán siendo tan escasos como la potencia (Sedlmeier y Gigerenzer, 1989), algo que Cohen (1962) ya criticó décadas antes. Sin embargo, nadie lo notará. Dentro de la estadística, a menudo se pasa por alto que existen enfoques bastante diferentes de las situaciones inciertas (Gigerenzer y Gaissmaier, 2011). Entre ellas se incluyen la replicación (por ejemplo, Nosek, Spies y Motyl, 2012), la minimización del error de medición, el análisis exploratorio de datos según Tukey (1977, véase el capítulo 5), los experimentos cuidadosamente diseñados para demostrar hechos y relaciones causa-efecto (por ejemplo, B.F. Skinner y sus legendarios experimentos con palomas), así como la heurística "rápida y frugal" (Reimer y Rieskamp, 2007) o los árboles de clasificación o regresión para la clasificación (Breimann, Friedman, Olshen y Stone, 1993), por nombrar sólo algunas de las variantes más comunes.

4.2 Esbozo de las teorías estadísticas

La estadística comienza con el concepto de probabilidad. Se podría decir que esto separa el trigo de la paja. Pero no es tan sencillo. Porque cada parte tiene sus argumentos.

4.2.1 El principio de probabilidad

La estadística se define por la forma en que se entiende y aplica el concepto de probabilidad. Pueden distinguirse tres variantes (Gigerenzer &Marewski 2015, p.430).

1. *Frecuencia relativa de una característica* a largo plazo: por ejemplo, las tasas de mortalidad, los casos de enfermedad, los accidentes o los retrasos a los que tiene que hacer frente cada día el ferrocarril.
2. *Propensión (física) de una característica*: por ejemplo, una moneda, una pelota, la disposición de un tablero de Galton.
3. *Probabilidad subjetiva*: se asigna a un suceso por diversas razones, por ejemplo, el conocimiento de expertos, el conocimiento previo basado en estudios e investigaciones bibliográficas, la evaluación de la credibilidad de los testigos y su capacidad para recordar ante un tribunal, o la sensación matinal de si hará un día soleado o lluvioso.

Las definiciones de *frecuencia relativa* y *propensión de una característica* pueden hacerse operativas con un esfuerzo razonable. En cambio, la definición de probabilidad subjetiva conduce al terreno de la incertidumbre y la gran variabilidad. Se trata de dos enfoques distintos de un problema, cada uno de ellos con información diferente. Por tanto, no se trata simplemente de que un formato codifique la información del otro de forma diferente. Aunque se puede interpretar una frecuencia relativa como una probabilidad, ésta no es cualitativamente la misma que la probabilidad que se puede expresar sin una frecuencia relativa. Más bien, son la propia información y los problemas asociados a ella los que son cualitativamente distintos, pero también se solapan a través de las matemáticas. El problema en sí no es el mismo. Es diferente si se trata de probabilidades o de frecuencias. Aunque el resultado matemático sea el mismo número, las interpretaciones no son las mismas.

Un ejemplo para aclararnos: ¿va a llover mañana? Esto puede responderse en términos bayesianos o frecuentistas. El planteamiento muestra que, a pesar de un resultado posiblemente idéntico, el camino es muy diferente:

• Cualquiera puede responder a la pregunta de si lloverá mañana mirando al cielo o teniendo en cuenta el tiempo de hoy, la previsión meteorológica, etc. Pero si pregunta con exactitud, descubrirá que mañana lloverá. Pero si pregunta exactamente, difícilmente obtendrá una respuesta del 100% con la máxima certeza, sino una aproximada. Si vivimos en una zona lluviosa, por ejemplo en el Amazonas o en el ecuador, la lluvia parece un acontecimiento bastante seguro para mañana, pero podríamos estar equivocados. Lo mismo ocurre con la época del monzón. Así pues, se mire por donde se mire, siempre hay lugar para un mínimo de incertidumbre y para un cambio inesperado del tiempo. Esto corresponde a la comprensión bayesiana de una probabilidad e independientemente de si ésta corresponde a la visión subjetiva del cielo o al complejo cálculo con los datos meteorológicos actuales. La experiencia pasada lleva a un aprendizaje, pero esto no es absoluto. La Antártida, por ejemplo, se considera un lugar frío, pero en el curso del cambio climático ha habido temperaturas de más de 15 grados centígrados incluso allí. El Sáhara era una sabana fértil al final de la última glaciación, hace entre 8.000 y 6.000 años, y hace muchos millones de años la Tierra también estaba completamente cubierta de hielo, incluso en el ecuador. El sol saldrá mañana para toda la vida, aunque esté cubierto de nubes. Pero si se infla hasta convertirse en una gigante roja dentro de unos miles de millones, eso podría cambiar en principio. La Tierra será entonces un planeta de lava incandescente. Los tiempos cambian y aprender de la experiencia, hablando en bávaro, siempre deja un poco de margen para desarrollos inesperados a corto plazo o incluso esperados a largo plazo. Si no existiera este espacio de incertidumbre, tendríamos que ocuparnos de ciertos acontecimientos de los que no hay nada que aprender y que tampoco es necesario calcular.

• A la inversa, se podría contar qué días del año ha llovido en un lugar determinado en los últimos años, o ha brillado el sol, o ha caído una tormenta, etc. A partir de estos recuentos (frecuencias), se pueden hacer deducciones frecuentistas y bastante complejas sobre cómo evolucionará el tiempo, hacer predicciones sobre el mañana, etc. Esto corresponde a la concepción frecuentista de la probabilidad. Se toma una muestra, se recogen los datos y se evalúan en condiciones asintóticas en la medida de lo posible – casi en vista del infinito – para poder aplicar los mejores algoritmos y supuestos de distribución disponibles. Esto impone condiciones mínimas pero importantes sobre el tamaño de la muestra y los

supuestos de distribución realizados. En el peor de los casos, las violaciones de estas condiciones previas significan que ya no es posible utilizar los métodos elegidos, por lo que tenemos que cambiar a otros métodos más robustos, que entonces responden a otras preguntas.

Si volvemos al punto de vista bayesiana, podemos hacer predicciones al cabo de un día, por así decirlo -es decir, basándonos en acontecimientos singulares-, pero a riesgo de cometer enormes errores a la vista de esta pequeña cantidad de datos. Nuestras conclusiones seguirían siendo coherentes, porque utilizamos los conocimientos previos de que disponemos. Nuestro conocimiento previo consiste en la ubicación geográfica, nuestra altitud sobre el nivel del mar, la época del año y otras características del lugar, como montañas que refrescan, una llanura, la proximidad del agua, etc.

Así, podríamos hacer predicciones casi sin datos empíricos, es decir, si no disponemos de nada. En cambio, en el caso de sólo N = 1, la estadística frecuentista tiene dificultades, ya que aún no corresponde a ninguna frecuencia real y, por tanto, los procedimientos aún no funcionan.

4.2.1.1 *Expectativas y subjetividad de las probabilidades*

Para entenderlo, veamos el teorema de Bayes:

$$Posterior = \frac{\Pr ior * Likelihood}{TotalEvidence} \qquad (4.1)$$

El teorema de Bayes establece que el estado actual del conocimiento (posterior) tras revisar los datos empíricos resulta del producto del conocimiento previo (Prior) y los datos (Likelihood) dividido por el número total de estados. El denominador es una cantidad que suele ser difícil en la práctica y que sólo puede calcularse mediante simulación, pero que en última instancia sólo normaliza el numerador. El numerador es relevante. Simplificado, el conocimiento surge así del conocimiento previo multiplicado por los datos empíricos. Si esto se repite, los datos empíricos suelen modelar la posterior casi por completo. Los conocimientos previos se obtienen mejor a partir de estudios anteriores sobre el mismo tema. Sin embargo, si esto falta, la información previa puede estar formada por conocimientos de expertos o incluso por suposiciones subjetivas.

Por tanto, es posible poner en forma matemática conocimientos existentes o incluso suposiciones subjetivas para incorporarlos al teorema de Bayes como conocimiento previo. Como se ha descrito, esto resulta ventajoso si no existen estudios preliminares u otros resultados empíricos sobre un objeto de investigación. Si existen, deben integrarse como conocimiento previo y las suposiciones subjetivas o el conocimiento experto profesional dejan de tener importancia, ya que los resultados empíricos suelen estar por encima de ellos.

> Recordatorio 4.1: Influencia de los conocimientos previos
>
> La cuestión central es si se permite que el conocimiento previo fluya en las ecuaciones estadísticas y qué formas puede adoptar este conocimiento.

El concepto de probabilidad es, pues, un asunto delicado. Si nos fijamos en el teorema de Bayes (véase el capítulo 6.3.2.4 para un ejemplo empírico), la información previa desempeña un papel igual al de los datos recogidos y está multiplicativamente conectada a ellos. En este sentido, deben ser tratados con el mismo cuidado. Por información previa se entienden las expectativas basadas en los conocimientos acumulados en

el pasado (es decir, antes de la recopilación de datos) ante la incertidumbre de hasta qué punto siguen siendo válidos para un estudio actual.

Sin embargo, según Gigerenzer y Marewski (2015), el peligro de la estadística bayesiana es atribuir de forma bastante universal una probabilidad interpretada subjetivamente a todos los sucesos inciertos e inciertos. Lógicamente, existe el peligro de la arbitrariedad. La "característica" de Bayes, es decir, el enfoque en las probabilidades y la inclusión de conocimientos contextuales o estudios empíricos previos, se convertiría así en una desventaja. La cuestión de la subjetividad que se deriva de ello no está clara, por lo que la abordamos en un discurso más extenso (véase el capítulo 6.5). La subjetividad existe en todas partes. Sin embargo, parece más significativa la necesidad de los investigadores de generar análisis de datos semiautomáticos.

Para comprender mejor esta situación, es necesario pasar de la probabilidad a la información. En una situación en la que no existe información empírica, heurística razonada o procesos de pensamiento para tomar una decisión, el uso de probabilidades puramente subjetivas parece legítimo. En ese caso, se trata de la única información disponible que el sentido común puede utilizar para sacar conclusiones lógicas. Sin embargo, si la situación de los datos cambia debido al empirismo y a la evaluación de la información, entonces, según el teorema de Bayes, la información previa (por ejemplo, las probabilidades subjetivas) debería tener cada vez menos peso y el resultado del teorema debería estar determinado cada vez más por el empirismo (los datos). Sin embargo, hay buenas razones por las que no siempre es así y por las que la información previa puede tener una influencia sustancial si, como expectativa inicial, contradice fuertemente los datos empíricos y, por tanto, tiene un efecto duradero en la posterior. En la mayoría de los casos, sin embargo, los proyectos de investigación concretos no implican muestras enormes ni conocimientos previos estables, por lo que las variables a priori casi siempre desempeñan un papel importante. En pocas palabras, existen varios procedimientos para hacer frente a este problema. Estos incluyen comprobaciones predictivas posteriores (véase el capítulo 6.8.4.4 para un ejemplo empírico) según Gelman et al. (2004) para validar el modelo en datos nuevos o simulados. Otra sugerencia procede de Spiegelhalter (2004), a saber, examinar los datos bajo un prior muy pesimista o muy optimista y comparar los resultados entre sí.

Ahora la pregunta es si esta dependencia de los resultados respecto a los valores previos, es decir, las expectativas, es buena o mala. Aquí es donde difieren las opiniones, tanto entre frecuentistas y bayesianos como dentro de la propia estadística de Bayes. Dentro de la estadística de Bayes, hay (Berger, 2014; Chick, 2005) representantes de la subjetividad de las probabilidades (por ejemplo, de Finetti, Rubin, Lindley) y los del enfoque objetivo de Bayes, que se remonta a Simon Laplace e incluye a investigadores como Je?reys o Jaynes. Los primeros se caracterizan por elegir una distribución a priori en función de convicciones personales subjetivas, mientras que el enfoque objetivo no las formula "arbitrariamente subjetivas", sino empíricamente justificables. Existen submodalidades para su aplicación. Uno de ellos es el enfoque de la Entropía Máxima (véase el capítulo 6.14), favorecido por Jaynes (1983). Tanto los enfoques subjetivos como los objetivos intentan hacer frente a la incertidumbre real sobre los parámetros y los modelos. Sin embargo, se dan cuenta de ello en cada caso ante una comprensión diferente en términos de teoría científica, que luego también tiene un efecto matemático. En el extremo, el campo objetivo llega hasta representantes que estiman la información a priori directamente a partir de los datos empíricos (enfoque Bayes empírico, Robbins, 1956; Casella, 1985), lo que, sin embargo, ya no corresponde al enfoque bayesiano clásico.

Desde nuestro punto de vista, la cuestión no sería si las probabilidades subjetivas tienen sentido o no, sino cómo relacionamos la información disponible de la forma más inteligente posible con los nuevos datos empíricos y cómo mantenemos la coherencia en nuestros argumentos. ¿De qué información verificable (es decir, conocimientos) se dispone y de qué manera se integrarán los datos futuros? ¿Cómo cambia esto los resultados y las interpretaciones de los análisis? Si la información previa sigue desempeñando un papel, esta influencia debe tenerse en cuenta y siempre forma parte de la interpretación. Dado que el proceso de análisis no puede ser totalmente objetivo, sino que sigue un criterio de verdad relativa, el debate es el medio elegido para determinar todas las variables y evaluar su influencia en el resultado. Por lo tanto, consideramos que las denominadas probabilidades subjetivas bajo la condición de máxima incertidumbre son completamente legítimas para una prioridad, siempre que se modelen y discutan explícitamente como parte del proceso de análisis. La calidad de un análisis, por ejemplo, basado en una revisión bibliográfica exhaustiva (Gelman y Carlin, 2014), puede ser muy precisa y fundamentada y aumentar significativamente la calidad del análisis

general e incluso mostrar sus limitaciones. Prescindir de él sería una negligencia. A falta incluso de esos datos, puede recurrirse al sentido común y a una reflexión cuidadosa para llegar a un estado de conocimiento previo. Este proceso puede hacerse explícito y comprensible. Lo que no nos parece bien son las decisiones sin justificaciones comprensibles, aunque resulten correctas a posteriori. Pero no cumplen los criterios del trabajo científico. El praradigma cualitativo también proporciona herramientas suficientes para llevar a cabo cuidadosamente este proceso de agregación de conocimientos cualitativos. No obstante, esto significa que se introduce un cierto grado de imprecisión e incertidumbre en el modelo global. Sin embargo, esto siempre se aplica a la inclusión de los parámetros del modelo.

4.2.2 Significado y finalidad de las estadísticas

La estadística consiste en ?descubrir cosas nuevas, confirmar o refutar expectativas, abordar la complejidad y representarla en modelos lo más sencillos posible. Podemos decirlo de forma aún más sencilla según Urban Studer (comunicación oral): la estadística es exploración de datos.

Una prueba estadística para confirmar una hipótesis requiere especificar con precisión lo que se va a probar. Como efecto secundario, al mismo tiempo es posible que se nuble la visión de las "cosas nuevas", ya que los datos no se examinan en busca de sus patrones implícitos, sino que se tratan en el marco de expectativas e hipótesis deductivas. Por tanto, la estadística, si sólo se aplica de forma con?rmatoria, sólo sirve condicionalmente para descubrir algo nuevo en los datos. Este caso se da como mucho cuando algo contradice claramente las expectativas y se buscan razones. De lo contrario, es probable que la gente tienda a ver la confirmación más que la refutación de sus suposiciones. Descubrir algo nuevo, sin embargo, requiere el uso de métodos abductivos o inductivos (véanse los cap. 2.3 o 2.1), como pedía Fisher (1935/1973). Fisher sugirió que sólo se realizara una prueba estadística cuando la situación no estuviera clara y se supiera poco sobre una zona objeto. Mucho más importante es un buen diseño de la investigación y un buen proceso de reflexión.

Así pues, la confirmación sólo es adecuada para poner a prueba las presuposiciones existentes. Si esto tiene sentido en casos concretos debe decidirse por separado en cada caso y depende de la cuestión de fondo y, sobre todo, de la interpretación que se aplique al concepto de probabilidad. Además, está la cuestión de si el conocimiento contextual debe y puede incluirse (por ejemplo, en la estadística bayesiana) o si se rechaza por no ser científico (entonces, estadística clásica). Las distinciones que se hacen entonces son de naturaleza cualitativa y, por consiguiente, tienen implicaciones para las matemáticas.

Mientras que la estadística clásica suele atribuirse al campo de la deducción (véase, por ejemplo, el diseño del estudio Neyman-Pearson), la estadística bayesiana suele atribuirse al campo inductivo. Sin embargo, esta distinción tan radical es anticuada e incorrecta, ya que es el diseño de la investigación el que determina la lógica subyacente del diseño y, por tanto, la importancia de los resultados. No es el tipo de análisis de los datos el que decide sobre la inducción o la deducción, sino que ésta representa en su selección una consecuencia del diseño subyacente. Sin embargo, esto no significa que todos los enfoques estadísticos sean adecuados para todos los diseños.

El propio Fisher (véase el capítulo 4.3.2) consideraba su planteamiento como una inferencia inductiva para descubrir algo nuevo en caso de escaso o nulo conocimiento en un campo nuevo a través de sucesos raros (basados en la significancia estadística o en valores p exactos; véase el capítulo 4.3.8 para las distintas variantes de uso del valor p). Mientras que el enfoque de Neyman-Pearson (véase el capítulo 4.3.3) tiene lugar en un marco hipotético-deductivo (palabra clave: análisis de potencia a priori), la inferencia es básicamente inductiva, ya que la población se infiere a partir de una muestra, que por cierto pertenece a los definidos, como critican Gigerenzer y Marewski (2015) en muchos estudios psicológicos (véase también la crítica al estudio de Gigerenzer y Marewski a la crítica del estudio de Bem, 2011a, cap. 4.4.2 y 6.8.1.6). Ya sea según Fisher o según Neyman-Pearson, los resultados de la estadística clásica, que se basan en la prueba de significancia, representan un procedimiento realmente inductivo dentro de un marco mayoritariamente deductivo.

A la inversa, no es cierto que la estadística de Bayes sea "sólo" inductiva. Por el contrario, los modelos y predicciones pueden y deben derivarse y probarse específicamente (Gelman & Shalizi, 2013), lo que corresponde a un enfoque deductivo. Ambas formas de estadística son, en principio, adecuadas para todas las tareas científicas si se tienen en cuenta cuidadosamente las características de los respectivos enfoques. Esto demuestra que no tiene sentido pensar constantemente en inductivo o deductivo. Más bien, la ciencia debe verse como un ciclo interminable de reflexión, exploración y revisión crítica (véase la Fig. 2.1, p.8). Es importante saber en qué fase se está trabajando actualmente y qué planificación requiere la siguiente fase para su realización.

En cuanto a la integración de CUAN/ CUAL, AED y las estadísticas descriptivas desempeñan un papel importante. Por ello, el AED se trata por separado en este libro (véase el capítulo 5). El AED se basa en la estadística descriptiva y sus métodos, pero debe entenderse como un enfoque independiente únicamente por la actitud de "descubrir en lugar de describir". Esto debe distinguirse claramente de la estadÌstica descriptiva en la aplicación de procedimientos confirmatorios (estadística inferencial), que sólo se ocupa de verificar las condiciones previas (por ejemplo, la distribución normal de los datos y los residuos) y las orientaciones (por ejemplo, los valores medios y las varianzas) antes de aplicar procedimientos de prueba estadística confirmatorios relativamente a ciegas.

Hoy en día, cuando se termina un estudio empírico, todos los enfoques coinciden en que debe repetirse para garantizar la constancia de los resultados. Difícilmente puede sobrestimarse el papel de la replicación para adquirir conocimientos, aunque esta necesidad dista mucho de reflejarse en la práctica de publicación pertinente de las revistas habituales. Una excepción notable es el artículo de Murre y Dros (2015). Los autores reproducen las curvas de memoria de Ebbinghaus. Una discusión crítica de las Hoffnungen puestas en las réplicas es llevada a cabo por Gigerenzer (2018).

La limitada utilidad de la estadística queda patente en los trabajos experimentales de B.F. Skinner. Skinner basó los resultados de sus investigaciones de forma experimental y no estadística (Skinner, 1948). También prescindió de muestras más grandes para poder llevar a cabo con precisión sus experimentos con palomas, cuidadosamente construidos y reproducibles. De este modo, el control de las condiciones adquiere un significado más profundo. En el transcurso de este proceso, se fundó una revista independiente, el Journal of the Experimental Analysis of Behaviour, para poder publicar con independencia de los editores que se dedicaban a la comprobación de hipótesis nulas (Gigerenzer, 1993, p.313).

> „Skinner continued to investigate one or a few pigeons under well-controlled conditions, rather than run 20 or more pigeons under necessarily less well-controlled conditions to obtain a precise estimate for the error variance.“

Así pues, se hizo hincapié en un control experimental estricto de las condiciones y en minimizar el error de medición, en lugar de en muestras grandes y en medir el error de medición después del estudio, cuando ya es demasiado tarde para hacer correcciones.

Si nos fijamos en los principales y aún válidos enfoques teóricos de la psicología (Pavlov, Freud, Piaget, Reich, Skinner, Miller, Köhler, Erikson, Kohlberg y muchos más), todos ellos no se basan en grandes muestras con todas las grandes evaluaciones estadísticas, y algunos de ellos no se basan en la estadística en absoluto. Más bien, la curiosidad y la creatividad, la constancia y la persistencia, así como la observación de la realidad, parecen haber sido factores decisivos. Y éstas son precisamente las personas y las teorías que siguen constituyendo los cimientos de la psicología y que deben figurar en todos los cursos de licenciatura. Quizá la tendencia a la estadística sea en parte responsable de que en la psicología actual apenas existan verdaderas teorías, sino, en el mejor de los casos, fragmentos de teoría.

4.2.3 Estadísticas no concluyentes

El análisis estadístico de los datos incluye, en primer lugar:

- *Estadísticas descriptivas*, que se limitan a describir datos.
- *Análisis exploratorio de datos* (detallado en el capítulo 5) según John W. Tukey (1915-2000), cuyo objetivo es la identificación de patrones y, principalmente, la perspicacia exploratoria.

4.2.3.1 *Estadísticas descriptivas*

Las estadísticas descriptivas pueden añadirse a las estadísticas clásicas y bayesianas. Las estadísticas descriptivas "sólo" describen los datos y no derivan de ellos inferencias, estimaciones, decisiones de prueba, etc. La estadística descriptiva calcula las características de los datos y las distribuciones, como la media, la mediana, los valores extremos, la desviación típica y la varianza, etc. Los datos también se pueden mostrar gráficamente con ayuda de una representación gráfica. Los datos también pueden describirse gráficamente con ayuda de gráficos.

En el contexto de la estadística frecuentista, la estadística descriptiva a menudo sólo servía para describir los datos antes de que siguieran los análisis frecuentistas "relevantes" propiamente dichos. Sin embargo, esto no tenía un propósito de descubrimiento exploratorio, como sugería Tukey (1977), por ejemplo. Los nuevos tiempos han cambiado esto. Las publicaciones actuales suelen esperar estadísticas descriptivas sobre las distribuciones relevantes de los datos para poder obtener una impresión sin inferencias posteriores. Además del tamaño de las muestras, suele incluir valores medios, desviaciones típicas y, posiblemente, estadísticas robustas.

4.2.3.2 *Estadísticas exploratorias*

Sin embargo, la estadística descriptiva se utiliza de forma inteligente y creativa en el contexto del análisis exploratorio de datos (AED, Tukey, 1962, 1977, 1980). El objetivo de EDA es identificar estructuras en los datos. Por lo tanto, exige mucho de los analizados y de su capacidad para encontrar patrones o cambiar de perspectiva para descubrir algo nuevo. Lógicamente, los datos examinados con EDA no deben utilizarse al mismo tiempo para pruebas (confirmación) si se trata de la misma pregunta. En EDA, los métodos gráficos de visualización de datos y las transformaciones (no) lineales de datos se utilizan sobre todo para poder destacar claramente diferencias, correlaciones, tendencias y similares. Curiosamente, la tendencia en la comprobación (por ejemplo, modelfitting, análisis de residuos, identificación de valores atípicos) de modelos complejos (por ejemplo, regresión/anova, HLMs/ MLMs, ...) también apunta cada vez más hacia el uso de métodos gráficos (Gelman & Shalizi, 2012; Kruschke, 2013c; Loy, Follett &Hofmann, 2016; Loy, Hofmann & Cook, 2016) en lugar de basar los argumentos únicamente en coeficientes.

4.2.4 Formas de estadística inferencial

Existen dos formas de estadística inferencial que, desde el punto de vista de sus respectivos defensores, a menudo tienen poco que ver entre sí (Jaynes, 2003, isb. cap. 16 para los antecedentes históricos). Difieren fundamentalmente en sus respectivas probabilidades.

1. *Estadística bayesiana* (véase el capítulo 6), basada en el teorema de Bayes, llamado así por el reverendo Thomas Bayes (1701-1761), pero cuya elaboración básica se remonta a Pierre-Simon Laplace (1749-1827).

2. *Estadística clásica frecuentista* (véase el capítulo 4) tras las dos teorías de Ronald A. Fisher (1890-1962) y Jerzy Neyman (1894-1981) y Egon Pearson (1895-1980), respectivamente, incompatibles desde el punto de vista de sus fundadores.

4.2.4.1 *Estadística frecuentista*

La estadística frecuentista, a menudo denominada estadística "de recuento" o "clásica" (aunque la de Bayes es en realidad más antigua), utiliza una definición de la probabilidad basada en sucesos contables (= frecuencias relativas, véase más arriba). Se centra en la probabilidad condicional de los datos dada una hipótesis denominada (nula) H_0 para rechazarla o no, en pocas palabras: $p(D \mid H_0)$. Así, las decisiones se basan en los datos y todas las conclusiones se refieren a la probabilidad de los datos y no a la de las teorías o hipótesis. Las muestras – puesto que los datos pueden fluctuar – se consideran manifestaciones del azar y el objetivo es eliminar el azar mediante la repetición (replicación). En la práctica común, la teoría de Fisher se mezcla de forma inadmisible con la de Neyman-Pearson y los coeficientes correspondientes se lanzan juntos de forma irreflectada (Hubbard, 2004). El crecimiento del conocimiento se define por la rareza de los acontecimientos bajo la validez de dicha hipótesis nula. La referencia (= hipótesis nula H_0) para determinar la rareza no se suele definir en términos concretos, sino que es una hipótesis denominada NIL (Gigerenzer, 2004b). La validez de la hipótesis nula H_0 no puede demostrarse mediante una prueba estadística clásica. Sólo puede rechazarse a un cierto nivel de significancia, pero nunca probarse. El objetivo es la generalización de los resultados a poblaciones infinitas y la identificación de los parámetros "verdaderos", que se suponen válidos en condiciones infinitas y que pueden estimarse a partir de los datos. Se aplica la teoría clásica del error (Lienert y Raatz, 1998) consistente en

Observación = valor real + error (de medición)

a realizar. En función de la teoría estadística clásica, la situación se complica. La teoría de Neyman-Pearson se ocupa de un ajuste cuidadoso del tamaño de la muestra, la fuerza del Efecto, la potencia y el nivel de significancia elegido, mientras que la de Fisher sólo se ocupa del valor p.

4.2.4.2 *Estadística de Bayes*

La estadística bayesiana, por su parte, trabaja con probabilidades condicionales de sucesos dados los datos y no las descuenta. Permite incluir conocimientos previos (inter)subjetivos o conocidos sobre un área para llegar a conclusiones coherentes no sólo con muestras muy pequeñas. La aportación de este conocimiento previo disminuye a medida que aumentan los datos y adquiere relevancia cuando se trata de extraer conclusiones plausibles ante información incompleta (es decir, falta de datos empíricos) para tomar decisiones (Studer, 1996b, 1998). La cuestión de la subjetividad o la objetividad, como ya se ha mencionado anteriormente, dista mucho de estar resuelta en el seno de la estadística bayesiana e implica posturas que pueden no estar integradas en absoluto. Por ejemplo, hay representantes que destacan el elemento subjetivo de la convicción personal como relevante (entre otros, vÈanse los capítulos 6.3.1 o 6.5.2, pero véase un punto de vista sobre la estadística clásica, capítulo 4.3.7), mientras que otros conceden importancia a las "comprobaciones predictivas posteriores / posterior predictive checks" (véase el capítulo 6.8.4.3) para evaluar la calidad predictiva de los modelos equipados (fitted models) en situaciones y datos nuevos. Se suele recurrir a la simulación o, de forma más elaborada pero claramente más informativa, a la replicación. Sin embargo, en general esto se practica demasiado poco en la ciencia. A diferencia de la estadística clásica, la estadística de Bayes puede aprenderse a partir de la experiencia. Técnicamente, se trata de completar el teorema de Bayes basado en el conocimiento existente (=conocimiento previo) con los nuevos datos y actualizar así el estado del conocimiento.

Además, existe el enfoque Bayes empírico, en el que se estima la información a priori directamente a partir de los datos. En el enfoque clásico de la estadística de Bayes, la información a priori siempre se conoce antes de recoger los datos, de ahí el nombre de información a priori. Utilizar esta información dos veces, como se hace en el enfoque Bayes empírico, parece cuestionable.

Asimismo, la estadística de Bayes conoce muchas decisiones (por ejemplo, en el contexto de los factores de Bayes, sobre las distribuciones a priori, etc.) que no siempre parecen estar justificadas desde el punto de vista sustantivo, sino que parecen surgir de una mezcla de convención, habitus y convicción. Esto se parece mucho a la elección igualmente infundada de los límites de significancia en el contexto de la estadística frecuentista. El problema fundamental de definir los criterios adecuados, a partir de los cuales se toma una decisión basada en consideraciones de fondo ("puede", "debe", "debería") y no por convención de la "comunidad científica" parece ser un arte y, contrariamente a lo que se supone, tiene poco en común con el enfoque estadístico per se. La existencia de corrientes tan diversas puede entenderse como una manifestación ejemplar de la "inconmensurabilidad de los paradigmas" dentro de un paradigma mayor (Kuhn, 1973; sobre las principales corrientes en la filosofía de la ciencia, cap. 3).

La traducción del conocimiento cualitativo a priori en una distribución matemática a priori se considera un arte entre los estadísticos bayesianos. Sin embargo, representa un ejemplo de integración satisfactoria de métodos: *mapear con precisión y expresar numéricamente el conocimiento cualitativo*. La estadística de Bayes examina la probabilidad condicional de las hipótesis *H* dados los datos empíricamente disponibles *D* o la información *I*, abreviado *p(H | D)* o *p(H | I)*. El procedimiento se justifica por el teorema de Bayes, que combina multiplicativamente la información previa con los datos recogidos para modelizar el estado actual de los conocimientos. El objetivo es *una solución relacionada con problemas concretos en condiciones finitas*. Nunca se trata del infinito, como presupone la estadística clásica. El conocimiento se acumula a través de nueva información añadida que puede compensarse directamente con el conocimiento existente. Así, los conocimientos existentes fluyen hacia los cálculos actuales. En la siguiente investigación, los resultados del empirismo anterior pueden incluirse como nueva información a priori en las ecuaciones (véase en el capítulo 6.15.2 un estudio de caso sobre los índices de aprobados en el tratamiento de drogodependencias). En principio, este proceso de acumulación de conocimientos puede continuar indefinidamente. Los parámetros se consideran variables aleatorias con una distribución de probabilidad asociada dados los datos manifiestos disponibles. En cambio, en la estadística clásica, los datos se consideran aleatorios y los parámetros fijos. La única cuestión aquí es si los profesionales bayesianos funcionan realmente de este modo, ya que implica replicar realmente los estudios. Dudamos que se trate de una práctica habitual, de nuevo con independencia del enfoque estadístico practicado. Por supuesto, replicar estudios no es trivial. Por un lado, hay que aprender de los errores del estudio existente conocido, lo que implica un cambio y un mayor desarrollo. Por otro lado, la replicación debe ser lo más exacta posible, pero esto sólo funciona si el nuevo estudio realizado es (todavía) comparable con uno anterior, lo que es directamente contrario a la exigencia de cambio y aprendizaje de los errores cometidos anteriormente. No existen criterios claros al respecto. Los límites parecen difusos. En este sentido, la replicación nunca es un asunto trivial. Para colmo, las réplicas se caracterizan por la falta de innovación y las revistas se toman demasiado en serio a sí mismas como para considerar útil publicar réplicas de forma amplia y continuada. Sin embargo, la literatura reciente muestra la importancia que se le ha dado al tema (por ejemplo, Ulrich, Erdfelder, Deutsch, Strauß, Brüggemann, Hannover, Tuschen-Caffier, Kirschbaum, Blickle, Möller & Rief, 2016). También hay ahora un número especial de *Psychologische Rundschau* (véase también Erdfelder & Ullrich, 2018; Fiedler, 2018) sobre el tema de la replicabilidad. Sólo los estudios que han sido reproducidos (varias veces) y realizados según estrictos criterios de investigación deben incluirse en los libros de texto como conocimientos sólidos. Dónde está ya el caso en las ciencias sociales – aparte quizás del experimento de Milgram, que fue replicado por Doli*n*ski, Grzyb, Folwarczny, Grzyba*ł*a, Krzyszycha, Martynowska y Trojanowski (2017), pero véase Perry (2013) para la relativización de la validez de los resultados originales.

4.2.5 La prueba estadística

Para empezar, subrayamos que la prueba estadística es una continuación directa del hilo rojo introducido al principio con las diferentes begri de probabilidad (secciones 4.2.1 y 4.3.9 respectivamente). Una prueba se basa en una probabilidad, se formule como se formule. Así pues, todas las conclusiones e interpretaciones dependen directamente del concepto de probabilidad: ¿se cuenta (frecuentista) o no?

Las conclusiones científicas en el ámbito cuantitativo numérico se basan – desde un punto de vista descriptivo – en la comparación de hipótesis (por ejemplo, predicciones, explicaciones) con datos observacionales. Si las hipótesis son correctas, esperamos que los datos de observación se ajusten a ellas; y si las hipótesis resultan no ser válidas y, por tanto, el modelo supuesto descalificado, los datos de observación deberían mostrar grandes desviaciones con respecto a estas expectativas. Entonces es necesario revisar el modelo, la teoría, la hipótesis, etc. y realizar nuevos estudios o replicarlos. Esta comparación entre los datos esperados y los observados constituye la base fundamental sobre la que se asientan las pruebas estadísticas, las pruebas de significación. La magnitud de la desviación de la hipótesis frente a los datos observados para decidir si rechazar o aceptar o no una hipótesis no es una cantidad eterna e independiente del tiempo, porque no se basa en leyes naturales, sino sobre todo en convenciones insensibles al contexto. En la estadística isb. frecuentista se han establecido durante décadas convenciones que ocultan el hecho de que determinar la discrepancia entre el modelo, la teoría, la hipótesis, por un lado, y los datos observacionales, por otro, es un problema no trivial y muy complejo para el que no existe una respuesta general, sino, en el mejor de los casos, una aproximación a lo largo de los respectivos requisitos de contenido. El problema existe igualmente en la estadística bayesiana, y algunos científicos intentan establecer en este campo convenciones que introduzcan determinados tamaños mÌnimos (palabra clave: factor de Bayes), por ejemplo, para identificar de forma semiautomática los hallazgos significativos. En resumen: la gente tiende a anteponer los valores numéricos a las consideraciones de contenido.

Técnicamente, entonces, existe una hipótesis *H* sobre un fenómeno *P* , que está siendo observado. El enfoque frecuentista clásico, basado en la hipótesis *H*, suele preguntarse cuál es la probabilidad de obtener determinados datos, por ejemplo, una determinada secuencia al extraer bolas de colores de una urna. Detrás de ello está la presunción de una conexión lógica o incluso causal entre una hipótesis específica y los posibles datos observables. La propia decisión de la prueba se basa en la probabilidad de los datos en relación con una hipótesis nula a menudo no especificada. El modelo M y la hipótesis se consideran verdaderos y conocidos, respectivamente, y los datos se consideran productos aleatorios. Además, la corrección de una hipótesis está sesgada por un error de medición que siempre está presente en contextos reales. Siempre se trata de saber si los datos se ajustan o no a la hipótesis, lo que en cada caso conlleva las correspondientes consecuencias para las decisiones sobre la acción. Este procedimiento común también se considera el modelo de mejores prácticas en las ciencias sociales. Jaynes (2003, p.84f.) contradice este procedimiento y establece el enfoque del razonamiento científico de forma bastante diferente al establecer los datos como verdaderos o conocidos y la hipótesis o varias hipótesis (H_1, H_2, . . ., H_n) etc. como desconocidas y, por tanto, como variables aleatorias para las que se pueden calcular probabilidades:

> „In virtually all real problems of scientific inference we are in just the opposite situation: the data *D* are known but the correct hypothesis *H* is not. Then the problem facing the scientist is of the inverse type: Given the data *D*, what is the probability that some specified hypothesis *H* is true? [...] Indeed, the scientist's motivation for collecting data is usually to enable him to learn something about the phenomenon in this way."

Este punto de vista conduce inevitablemente a una nueva perspectiva de las hipótesis y los datos de observación, porque estas probabilidades no pueden calcularse de forma frecuentista. En particular, el conocimiento previo antes de observar los datos desempeña ahora un papel, ya que según Jaynes (ibid.) hay que preguntarse:

> „'What do you know about the hypothesis *H* after seeing the data *D*?' cannot have any defensible answer unless we take into account: 'What did you know about *H* before seeing *D*?' But this matter of previous knowledge did not figure in any of our sampling theory calculations. When we asked: 'What do you know about the data given the contents (*M, N*) of the urn?' we did not seem to consider: 'What did you know about the data before you knew (*M, N*)?'"

De este modo, Jaynes esboza con elegancia la discrepancia entre la estadística clásica (= probabilidad de los datos dada una hipótesis en su mayor parte inespecífica) y la estadística bayesiana (= probabilidad de las hipótesis dados los datos).

4.2.6 Métodos mixtos

El tipo de procedimiento de análisis de datos en cuestión no determina la decisión a favor o en contra de los métodos mixtos: las posibilidades de integrar métodos de investigación son demasiado diversas (véase el capítulo 13). La integración de métodos no sólo abarca el ámbito de la estadística y el análisis cualitativo de datos, sino todo el proceso de investigación. Si un determinado método analítico parece tener sentido en el contexto de una pregunta de investigación, debe utilizarse. En las siguientes explicaciones, no tratamos por separado los procedimientos específicos de análisis de datos. Esto es responsabilidad de los libros de texto mencionados.

A primera vista, algunos procedimientos de análisis de datos parecen, en efecto, sólo condicionalmente aptos para ser preseleccionados como *Métodos Mixtos*, si se trata precisamente de complementar los datos cualitativos con una perspectiva estadística. Entre ellos se encuentran la escala de Rasch (Rost, 2004) y los modelos de ecuaciones estructurales (SEM), muy populares en psicología y pedagogía empírica. En el caso de los SEM, en particular, a menudo parece que nunca se llega a nada y que su uso práctico concreto no est-claro porque no se conocen todas las demás configuraciones posibles, es decir, faltan las alternativas para poder apreciar en su conjunto la configuración que se presenta como la solución final frente a esas alternativas. Además, el procedimiento no es necesariamente sólido y la demostración de causalidad sin un diseño longitudinal apropiado a lo largo del tiempo puede no tener éxito en absoluto, incluso si el modelo es adecuado (Rietz, Rudinger & Andres, 1996). Aunque todo parece siempre bastante coherente en las publicaciones pertinentes, uno se pregunta cómo se desarrollan realmente los resultados en la realidad, para poder relacionar los resultados de los estudios con la realidad. Sin embargo, todo esto no sería un obstáculo para los Métodos Mixtos y puede discutirse con el material de datos. Especialmente con los SEM, vemos un gran obstáculo a la hora de situar el nivel de codificación, las interpretaciones del texto y las pruebas de hipótesis sobre el texto en un contexto razonable que siga la línea de la pregunta de investigación y no se convierta en un enfoque escopeta. Como ya se ha mencionado, no existe una justificación comprensible a nivel de las variables utilizadas de por qué debe darse preferencia a este modelo resultante frente a otro. No se trata de los criterios de calidad posteriores como AIC, BIC, etc. (véase el capítulo 6.8.2) – aunque es una lógica extraña con los SEM probar un modelo de forma clásica-estadística con la esperanza de que no sea rechazado y el conocimiento en este caso es precisamente cero. El conocimiento es nulo, porque el no rechazo de un modelo desde un punto de vista clásico-estadístico no conduce a ninguna ganancia de conocimiento. Por lo tanto, se utilizan en su lugar otros criterios de información (véase el capítulo 6.8.2), sin ignorar por completo la prueba clásica-estadística. Básicamente, se trata de la selección de variables y sus relaciones entre sí y de cómo se van a modelizar. Los datos cualitativos suelen carecer de criterios claros, como constructos definidos teóricamente, conglomerados, cuestionarios, etc. Por lo tanto, se trata de un proceso reconstructivo que, como explican Glaser y Strass (1998) en el contexto de la teoría fundamentada (véase el capítulo 9.3), tiene sus raíces en los propios datos. Merecería la pena investigar si un procedimiento cualitativo por etapas con una justificación precisa de la selección de variables y direcciones o relaciones entre variables podría resolver este problema. Hasta entonces, vemos con ojos muy críticos el uso de SEM en el contexto de los procesos de análisis reconstructivos y tenderíamos a desaconsejarlo. Además, el procedimiento se utiliza con el objetivo de la confirmación.

Por otra parte, el escalado de Rasch se encuentra sobre todo en el ámbito de la medición del rendimiento (por ejemplo, la medición de competencias) y no se practica de forma muy heurística debido a sus peculiaridades matemáticas, lejos de la idea de integración de métodos. Es probable que muchas cuestiones cualitativas de interés en las ciencias sociales no sean accesibles en absoluto mediante el escalado de Rasch o un análisis similar. Esto significa que esta clase de métodos estrechamente definidos, con sus elevadas exigencias sobre los datos y, por tanto, sobre el proceso de recopilación de datos, prácticamente desaparece. El análisis de las interacciones sociales, por ejemplo, no se corresponde con el análisis de las competencias matemáticas, sino que entra en un ámbito cualitativamente distinto con sus propias reglas y estructuras. Después de todo, no debería darse el caso de que el análisis de datos define el enfoque metodológico del fenómeno investigado sólo para poder aprovechar el proceso de análisis de datos. Eso sería una burda tontería.

La decisión *a favor de algo* es siempre una decisión *en contra de todo lo posible*. Nuestro mantra, que ya se ha expresado varias veces, es, por tanto, *que debe examinarse en cada caso concreto si un procedimiento es adecuado para el asunto de que se trate*. A continuación, nos centraremos en procedimientos comunes como los modelos lineales (generales), los análisis de tablas o las clasificaciones, que pueden servir muy bien en la integración de un método. Esto no excluye que el SEM y el escalado Rasch también puedan ampliar el repertorio de métodos mixtos en el futuro. En la actualidad, tienden a no poder hacerlo.

4.2.7 Ajustar las estadísticas

Las estadísticas se prestan a la manipulación, intencionada o no. En este sentido, se dice que Winston Churchill (1874-1965) dijo: "Las únicas estadásticas en las que se puede confiar son las que uno mismo falsifia". Sin embargo, los investigadores de Churchill no encuentran esta frase en ninguna de sus obras. No obstante, la frase tiene cierta validez, porque sólo el examen de los datos demuestra si una investigación contiene errores y de qué naturaleza son.

Dado que los datos empíricos son variables aleatorias, en la estadística clásica, a diferencia de la estadística bayesiana, los nuevos datos no pueden integrarse sin más en un sistema de conocimiento existente, por ejemplo, actualizando simplemente las fórmulas y sus resultados preliminares anteriores. Para muchos parámetros de un diseño, no existen reglas vinculantes ilimitadas en el tiempo. El tamaño de una muestra o su composición se responden de forma diferente según el punto de vista. En la estadística clásica, las muestras deben considerarse sucesos aleatorios y cualquier intervención en este supuesto sistema naturalmente aleatorio puede tener un efecto. Cada investigación supone una nueva coincidencia, lo que puede llevar a situaciones absurdas en investigaciones de mayor envergadura. Esto incluye la cuestión de dónde hacer el corte de una muestra aleatoria a la siguiente en la recogida de datos. Así, el cambio de corte de una muestra puede dar lugar a resultados completamente distintos. En casos extremos, esto puede llevar al llamado "p-hacking" o "fishing for significance" (Bermeitinger, Kaup, Kiesel, Koch, Kunde, Müsseler, Oberfeld-Twistel, Strobach & Ulrich, 2016), en el que se recopilan datos hasta que un resultado resulta es estadísticamente significativo y conforme a la hipótesis; y entonces se da por concluido el estudio (véase también Gelman & Loken, 2013; Lakens, 2015; Simmons, Nelson & Simonsohn, 2011; Simmonsohn, Nelson y Simmons, 2014; Simmons y Simonsohn, 2017). La entrada del blog y el debate posterior en Cross Validated (usuario Silverfish, 2016) ofrecen una introducción a la bibliografÌa y los conceptos relevantes para el tema.

Otra variante consiste en incluir parámetros en el modelo. En principio, es posible probar todas las variables entre sí y alguna comparación "ya resultará estadísticamente significativa", circunstancia que podría eliminarse en gran medida mediante un análisis de potencia previo según Neyman-Pearson (véase el capítulo 4.3.3) y comparaciones planificadas, si se practicara.

Otra posibilidad es "retocar" la composición de la muestra, eliminar los valores atípicos, aumentar el tamaño de la muestra, omitir condiciones, formar una media global sobre las variables dependientes y utilizar sólo ésta, etc. Los resultados pueden utilizarse para calcular la media. Krämer (2000) ha resumido algunas de las técnicas más sencillas para mentir con la estadística, lo que no implica que aquí asumamos en general que los investigadores falsifican intencionadamente sus datos y análisis. Se trata sin duda de excepciones. Simonsohn (2014) ofrece un ejemplo con código R que simula diferentes modos de p-hacking al tiempo que muestra que el mismo problema también implica factores de Bayes (véase la sección 6.8.1). Sin embargo, los errores metodológicos (Gelman & Loken, 2013) se encuentran a menudo en la necesidad inconsciente de los investigadores de encontrar resultados que se ajusten a la hipótesis, y recorren todo el diseño de la investigación, incluyendo la formulación de la hipótesis, la generación de la muestra y los instrumentos utilizados, y no necesariamente (sólo) a través de los análisis estadísticos. Todo el proceso se ve afectado, es decir, la generación y preparación de los datos que finalmente se analizan.

> „A naïve student of Bayesian inference might claim that because all inference is conditional on the observed data, it makes no di erence how those data were collected. [...] The essential flaw in the argument is that a complete definition of ‘the observed data’ should include information on how the observed values arose, and in many situa-

tions such information has a direct bearing on how these values should be interpreted" (Gelman, Carlin, Stern & Rubin, 2004, S.198)

Hay que distinguir si los científicos falsifican deliberada e intencionadamente sus investigaciones o simplemente caen en sus propias suposiciones implícitas. El primero es probablemente el caso menos frecuente. Es probable que la orbitación inconsciente de los diseños según las propias expectativas y, por tanto, en contra de la sugerencia de Popper de poner a prueba críticamente los propios edificios teóricos, ocurra con mucha más frecuencia y sin ser detectada. Sin embargo, este procedimiento no es correcto.

4.2.8 Otras lecturas y programas informáticos

Como libro de texto introductorio a la estadística clásica, recomendamos Eid, Gollwitzer y Schmitt (2010), que actualmente consideramos la referencia en el área germanófona. Para modelos de regresión y modelos jerárquicos o multinivel (HLMs/ MLMs), recomendamos Gelman y Hill (2007) y Pinheiro y Bates (2009). Lo que estos libros tienen en común es que todos hacen hincapié en la integración con R (R Development Core Team, 2019d), nuestro software libre preferido para el análisis estadístico de datos. R se considera actualmente la lingua franca de la estadística y ofrece paquetes especializados para cada área temática. En el sitio web de CRAN encontrará un resumen de los paquetes. Además, en la actualidad existen innumerables libros y guías prácticas o viñetas sobre paquetes de R para implementar la estadística directamente con R, así como una enorme colección de artículos originales a los que remitirse cuando sea necesario. Consideramos esencial la combinación de libro de texto y práctica con R o software comparable para profundizar en los conocimientos estadísticos. Con R, el aprendizaje de la estadística ha dado un giro significativo; el aprendizaje ya no consiste (solo) en fórmulas matemáticas, sino en código R (McElreath, 2015). Incluso los blogposts en internet alcanzan así un nivel muy alto – si, por ejemplo, proceden de expertos como Gelman (2019b), se discuten a nivel de pares y contienen código R verificable o reproducible. En nuestra opinión, esto legitima la citabilidad de estas fuentes de Internet para un trabajo científico serio. Aunque aquí falta sobre todo la revisión por pares (las excepciones son los debates de expertos directamente en la entrada del blog en cuestión), hemos observado en los últimos >10 años que la calidad de las entradas de blog y los artículos publicados en sitios web privados puede alcanzar un nivel excepcionalmente alto y bien fundamentado. También incluimos los artículos publicados en línea que han sido rechazados por las revistas. En principio, el rechazo de un artículo por parte de una revista no dice mucho sobre su calidad fundamental, sino en muchos casos sólo algo sobre la orientación del contenido de una revista y las expectativas o normas de lo que se acepta socialmente en la época correspondiente. El físico E.T. Jaynes, al que se cita varias veces en el libro, también vio cómo sus artículos bayesianos al principio no se publicaban y eran rechazados, y más tarde desempeñó un papel pionero en su campo. Esta incertidumbre se aplica al número de veces que se cita un artículo, pero esto puede ser un artefacto (Gigerenzer & Marewski, 2015) y ni siquiera dice nada sobre si la persona que cita un artículo lo ha leído y a la vez lo ha entendido y reflectado. La seriedad de la fuente respectiva debe evaluarse siempre individualmente, tanto si se publica en una revista como en un sitio web. En general, los buenos artículos en línea remiten a la bibliografía pertinente y no se formulan aislados de ella, sino que se insertan en los temas de debate habituales como los "artículos normales de revista" y hoy en día suelen ofrecer código R reproducible. Exigimos una bibliografía conforme a las normas habituales. Como científicos, por tanto, no debemos limitarnos a ciertas fuentes, sino pensar activamente junto a ellas y evaluar la seriedad de cada una por separado.

En cuanto a la estadística bayesiana, recomendamos los trabajos de Jaynes (2003), Gelman, Carlin, Stern y Rubin (2004), Gelman y Hill (2007), Gill (2008), Bolstad (2007), Gregory (2006) y Kruschke (2015b) y McElreath (2015). En el mundo germanoparlante, Tschirk (2014) ofrece una introducción fácil de entender que hace un excelente trabajo al recoger las diferencias con la estadística frecuentista y demostrarlas con ejemplos concretos. O'Hagan, Buck, Daneshkhah, Eiser, Garthwaite, Jenkinson, Oakley y Rakow (2006) ofrecen un libro sobre el tema de la generación de distribuciones de probabilidad a priori basadas en el conocimiento experto. Los innumerables artículos de revistas y blogs ya mencionados completan el panorama.

4.3 Estadísticas clásicas

4.3.1 Competencia en estadística clásica - Fisher frente a Neyman-Pearson

La estadística clásica-frecuentista se divide en los enfoques de Fisher y Neyman-Pearson. Estos se enzarzaron en una enemistad personal que duró años y que comenzó con la aversión mutua entre Fisher y Karl Pearson, el padre de Pearson (Jaynes, 2003). Tanto Fisher como Neyman-Pearson delimitaron claramente sus respectivas teorías (Lehmann, 1993). Resulta aún más sorprendente que más tarde y todavía en los libros de texto actuales pueda leerse una mezcla poco acertada de ambos enfoques. Los antecedentes pueden consultarse en Gigerenzer (1993, 2004b), Hubbard (2004) y Hubbard y Bayarri (2003). En general, los conceptos de Fisher se entrelazan con los de Neyman-Pearson sin una línea clara de razonamiento y se presentan como un supuesto todo. Esto no es así. Ambos son enfoques diferentes que, de hecho, pretenden cosas distintas y no aplican los mismos criterios al análisis de datos. Como resultado, se extraen diferentes interpretaciones y conclusiones de los mismos datos.

4.3.2 La teoría de R.A. Fisher - inferencia inductiva

La teoría de R.A. Fisher está diseñada para validar algo nuevo mediante pruebas estadísticas y, de este modo, adquirir conocimientos. Esto sigue la lógica del conocimiento inductivo, que Fisher defendió con vehemencia como fuente de progreso científico (Fisher, 1935/1973, p.3).

> „I have assumed, as the experimenter always does assume, that it is possible to draw valid inferences from the results of experimentation; that it is possible to argue from consequences to causes, from obersvations to hypotheses; as a statistician would say, from a sample to the population from which the sample was drawn, or, as a logician might put it, from the particular to the general."

Fisher introdujo los límites críticos del 1%, 5% y 0,01% para establecer la significación, que se utilizan hoy en día de forma de cabildero y sin reflexionar. Esto condujo posteriormente a la prueba de hipótesis nula no reflexionada (NHST), que, sin embargo, no fue formulada de esta manera por Fisher. La razón real para elegir estos valores fue que a Fisher no se le permitió utilizar las tablas de Karl Pearson, el padre de Egon Pearson, y en aquella época no existían ordenadores para calcular valores p exactos (Jaynes, 2003). Así pues, Fisher sólo disponía de unas pocas tablas para los valores p y las aprovechó al máximo. Sin embargo, en el debate con Neyman-Pearson, Fisher cambió de opinión sobre la aplicación rígida de las barreras de significancia y tendió a informar sólo de los valores p exactos (Gigerenzer, Krauss y Vitouch, 2004, p.11). La cantidad de citas de Fisher que se pueden encontrar en los libros de estadística y artículos de revistas pertinentes es relativamente fácil de sondear y a menudo se repite, es decir, distintos autores citan las mismas frases una y otra vez. Si se lee a Fisher en el original, surge un sentido claramente más complejo de sus afirmaciones, de modo que las citas relevantes parecen inadmisiblemente truncadas, como algo así como el límite de significación del 5%. Por lo tanto, nos tomamos la libertad de permitir que Fisher se exprese aquí con bastante más extensión de lo habitual. Con frecuencia se encuentran en la literatura las mismas citas breves de Fisher, que desgraciadamente distorsionan sus preocupaciones. Hacia el final de su vida, Fisher (1956/1973, p.45) advirtió contra el uso poco eficaz de pruebas de significancia con siempre los mismos valores de umbral crítico:

> „The attempts that have been made to explain the cogency of tests of signi?cance in scientific research, by reference to hypothetical frequencies of possible statements, based on them, being right or wrong, thus seem to miss the essential nature of such tests. A man who 'rejects' a hypothesis provisionally, as a matter of habitual practice, when the significance is at the 1 % level or higher, will certainly be mistaken in not more than 1 % of such decisions. For when the hypothesis is correct he will be mistaken in just 1 % of these cases, and when it is incorrect he will never be mistaken in rejection. This inequality statement can

> therefore be made. However, the calculation is absurdly academic, for in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas. Further, the calculation is based solely on a hypothesis, which, in the light of the evidence, is often not believed to be true at all, so that the actual probability of erroneous decision, supposing such a phrase to have any meaning, may be much less than the frequency specifying the level of significance. "

Fisher parte así de la base de que el mundo cambia y, por tanto, los distintos contextos requieren un enfoque diferente. Esto incluye cambiar los umbrales que determinan la significación estadística o no. Aparte de eso, no se entiende como única instancia para tomar decisiones. Más bien, los supuestos teóricos y las pruebas empíricas van de la mano para llegar a un juicio razonable. Así, los cálculos se sitúan bajo las hipótesis y, si suenan improbables, hay que ponerlos a prueba y no tomarlos a ciegas como el resultado final basado en una prueba estadística. Fisher se acerca así bastante al pensamiento bayesiano y biográficamente también se ocupó de este enfoque, ya que allí una hipótesis sigue siendo improbable a pesar de la evidencia empírica, por ejemplo, si la expectativa pegadiza reside en un nivel muy improbable u otros factores limitan la validez del empirismo. Y si se busca la redacción exacta y completa con la que Fisher discute la barrera del 5% de significación, suena mucho menos dogmática de lo que era y es comúnmente sostenida (Fisher, 1935/1973, p.13). Volvemos a dejar que Fisher hable por sí mismo. Las citas más largas de Fisher muestran que las citas muy cortas habituales de Fisher (por ejemplo, sobre el umbral de significación del 5%) tienen un fuerte efecto distorsionador y, por tanto, se comete una injusticia con Fisher en muchos aspectos. Fisher era un grupo de presión que propagó con fuerza su enfoque, pero no se le puede acusar de haberse ceñido estáticamente a su pensamiento, como demuestran en particular sus declaraciones tardías sobre el valor p.

> „It is obvious that an experiment would be useless of which no possible result would satisfy him [= 'the experimenter', nota de los autores]. Thus, if he wishes to ignore results having probabilities as high as 1 in 20 — the probabilities being of course reckoned from the hypothesis that the phenomenon to be demonstrated is in fact absent — then it would be useless for him to experiment with only 3 cups of tea of each kind. [...] It is usual and convenient for experimenters to take 5 per cent as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard, and, by this means, to eliminate from further discussion the greater part of the fluctuations which chance causes have introduced into their experimental results. No such selection can eliminate the whole of the possible effects of chance coincidence, and if we accept this convenient convention, and agree that an event which would occur by chance only once in 70 trials is decidedly 'signi?cant,' in the statistical sense, we thereby admit that no isolated experiment, however signi?cant in itself, can su?ce for the experimental demonstration of any natural phenomenon; for the 'one chance in a million' will undoubtedly occur, with no less and no more than its appropriate frequency, however surprise we may be that it should occur to us. In order to assert that a natural phenomenon in experimentally demonstrable we need, not an isolated record, but a reliable method of procedure. In relation to the test of signi?cance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result."

Si recapitulamos la cita, está claro que no se centra en la comprobación singular de la significación de estudios individuales. Por el contrario, Fisher hace hincapié tanto en la cuidadosa planificación y ejecución de los experimentos como en la repetición de los estudios, es decir, de tal manera que los efectos experimentales puedan demostrarse repetidamente y, por tanto, arrojen resultados significativos fiables. Además, queda claro que Fisher consideraba el famoso 5% como una convención y sin ninguna otra conexión más profunda con el objeto de estudio, lo que abre bastante espacio para seguir debatiendo. Sin embargo, el difunto Fisher no suena en absoluto dogmático, al contrario que las revistas y los científicos que lo aplicaron y que, obviamente, no lo entendieron.

Según Fisher, las pruebas de significación sólo desempeñan un papel cuando se sabe poco o nada sobre un área de investigación. Hay que hacer más hincapié en el buen diseño y su aplicación. Fisher se acerca más a Tukey y al AED (véase el capítulo 5) que Neyman-Pearson en cuanto a la idea, ya que se ocupa del conocimiento, que examina mediante una prueba en caso de información incompleta. Tukey (1977), por su parte, sigue un razonamiento cualitativo creativo de los datos cuantitativos y sus transformaciones. No se

puede culpar al difunto Fisher de que las hipótesis NIL estadísticas suelan formularse como hipótesis nulas. Fisher no lo recomendó. Jaynes (2003) sospecha incluso que Fisher sería hoy un bayesiano.

Sin embargo, la obra de toda una vida de Fisher no sólo incluye trabajos sobre biología, genética y estadística (por ejemplo, Anova, Maximum Likelihood), sino en general sobre diseño experimental y diseño de investigaciones (Fisher, 1925/1973, 1935/1973, 1956/1973). Esto incluye la aleatorización como base de la selección de muestras experimentales, que Jaynes (2003, p.497 y capítulo 10) explica:

> „Of course, Fisher's randomized planting methods — which we think to be not actually wrong, but hopelessly ineffcient in information handling — ... It appears to be a quite general principle that, whenever there is a randomized way of doing something, there is a nonrandomized way that delivers better performance but requires more thought."

Esta cita ataca un punto clave de la obra de Fisher, a saber, que la cognición corresponde a una desviación del azar y que las realidades del campo de investigación de Fisher, la biología, no son simplemente transferibles a cualquier otro campo. La aleatorización no pretende otra cosa, a saber, encontrar desviaciones sistemáticas del azar partiendo del supuesto de que las repeticiones infinitas eliminan toda aleatoriedad y sólo queda la sistemática subyacente. Jaynes (2003, cap. 17.7), por su parte, señala que no se trata de azar, sino de relaciones causa-efecto de la información que pueden reconstruirse con una investigación cuidadosa sin recurrir a condiciones artificiales de azar. Se trata de una estrategia cualitativa para reconstruir las relaciones causa-efecto de forma metodológicamente controlada, como se practica, por ejemplo, en el enfoque de la hermenéutica objetiva (véase el capítulo 11). Si, por el contrario, se operacionaliza la aleatoriedad, esto significa que un gran número de influencias actúan de forma no dirigida y deberían anularse mutuamente, mientras que las influencias de interés actúan de forma dirigida, a ser posible sin interacción con las "aleatorias". La aleatoriedad en este sentido no es el caos, sino la consideración simultánea de muchas entradas dirigidas de manera diferente con el fin de aislar a unos pocos en ellos – los e ectos a investigar, que deben ser uniformemente dirigidos. Esto no cuestiona el hecho de que detrás de cada "factor aleatorio" exista también un complejo causa-efecto. Una definición de azar es difícil de todos modos, aparte de las relaciones causa-efecto inespecíficas, y también es algorítmicamente irresoluble, ya que los números aleatorios en el ordenador dependen directamente de sus valores iniciales y, por tanto, pueden reproducirse exactamente (set.seed() en R). Por tanto, no es de extrañar que la estadística clásica pretenda trabajar en condiciones asintóticas en la medida de lo posible -es decir, comparables a tamaños muestrales infinitos- para poder sacar sus conclusiones de generalización. En muchos casos, esto no es erróneo – desde el punto de vista de la teoría de la información – sino desesperadamente ineficaz.

Fisher puede resumirse en que, en realidad, se trata de un buen diseño. Cuando se comprueba una hipótesis nula, el nivel de significación debe calcularse después del experimento e informarse con precisión, y todo esto sólo tiene sentido si se conoce poca información sobre el objeto de estudio. Las variables relevantes en Neyman-Pearson (potencia, tamaño de efecto, intervalos de confidencia, s. cap. 4.3.3.1 o 4.3.3.6) faltan completamente en Fisher. El tamaño de la muestra se calcula π por pulgada, si acaso. Además, está el lobby que Fisher utilizó para difundir sus ideas (Gigerenzer, Krauss & Vitouch, 2004; Gigerenzer, 2004b). El enfoque de Fisher es una herramienta útil en el arsenal estadístico en aquellas situaciones en las que no está claro qué está sucediendo realmente y ni siquiera existen hipótesis alternativas para especificar (Gigerenzer & Marewski,2015, p.434).

Por desgracia, sus interesantes y significativas reflexiones hacia el final de su vida no fueron suficientemente acogidas por la comunidad científica y apenas se difundieron.

4.3.2.1 Beber té y reconocer la leche - un experimento de ejemplo según Fisher

En su libro "The Design of Experiments" (1935/1973, p.11) Fisher describe un experimento con una mujer que afirma probar al beber té si primero se vertió leche en la taza y luego el té o, a la inversa, primero el té y luego la leche. La bibliografía identifica a esta mujer como la *Dra. Muriel Bristol*, una ficóloga (algóloga), es decir, una botánica especializada en algas (Salsburg, 2001). Para comprobar experimentalmente esta afirmación, Fisher creó un experimento. En ella, a Lady Bristol se le presentó una variante "té antes que leche"

cuatro veces y la otra variante "leche antes que té" cuatro veces. El orden de las muestras de té fue aleatorio. Lady Bristol recibió todos los detalles del experimento por adelantado. Su tarea consistió entonces en dividir las muestras de té en los dos grupos. Al llevar a cabo los detalles, Fisher utilizó este ejemplo para desarrollar la prueba exacta de Fisher que lleva su nombre (Fisher, 1935/1973, Cap. II), que se basa en la distribución hipergeométrica y, además, en la combinatoria y la permutación. Esto corresponde a "sacar *sin* volver a poner en su sitio", ya que Lady Bristol sabía de antemano lo que iba a pasar (por ejemplo, cuatro muestras de té con primero leche y luego té o viceversa). Así, la probabilidad podría determinarse mediante permutación, comparable al sorteo de los números de la lotería. Si a Lady Bristol se le hubiera ocultado cuántas variantes comprendía el experimento en cada caso, se podría haber utilizado la distribución binomial o la prueba binomial, ya que ésta presupone extracciones independientes. Esto corresponde al "sorteo con devolución". Con la prueba de Fisher, hay empates dependientes porque el número de las variantes respectivas se conoce de antemano.

Con n = 8 muestras de té, la asignación correcta del grupo de cuatro "leche antes del té en la taza" puede elegirse de 70 maneras o n!. Partimos del conjunto de todas las secuencias para ocho dividido por todas las posibilidades para cuatro asignaciones correctas, lo que apunta al coeficiente binomial

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$ (ptII_quan_classicstats_Fisher_ladyteataste.r)

```
> # n = 8
> # k = 4
> # n!/(k!*(n-k)!)
> (8*7*6*5*4*3*2*1) / (4*3*2*1)
[1] 1680
> # =
> factorial(8)/factorial(8-4)
[1] 1680
> # = 1680
```

Dado que esto no sólo contiene todas las posibles variantes de cuatro vías, sino al mismo tiempo todas las secuencias de estas posibles variantes de cuatro vías, las secuencias redundantes (n - k)! deben eliminarse, porque no son la cuestión aquí.

```
> # number of different sequences for the four group
> 4*3*2*1
[1] 24
>
> prod(4:1)
[1] 24
```

Ahora dividimos el conjunto de todas las variantes (= 1680) por el número de secuencias (= 24) o k!

```
> # all variants divided by different sequences of the four group
> 1680/24
[1] 70
> # = 70
>
> # 8-over-4 Binomialkoeffizient ->
> # possible ways to classify 4 of 8 cups properly as tea first or v.v.
> # binomial coefficient
> # possible ways to classify 4 of 8 cups of tea properly
> # with "tea first" (=four group)
> choose(8,4)
[1] 70
```

La probabilidad de acertar las cuatro variantes en 8 muestras de té es de una entre setenta,

```
> # chance to get all correct classified
> # in case no discrimination abilities are present
> # prob per cent
```

```
> 1/70
[1] 0.01428571
```

es decir, el 1,43%, por debajo del umbral del 5% preferido por el Fisher temprano. Por tanto, la prueba estadística de la hipótesis nula – asignar las muestras de té mejor que el azar – sería estadísticamente significativa. Esto puede determinarse directamente mediante factoriales:

```
> # small function just to calculate this
> teataste <- function(teeproben, positive)
+ {
+ return(factorial(teeproben)/
+ (factorial(positive)*factorial(teeproben-positive)))
+ }
> # call:
> teataste(8,4)
[1] 70
```

Tabla 4.1: Experimento de Fisher sobre el sabor del té

Filiaciones correctas		Combinaciones	p	cum. p	cum.p %
0	dhyper(0,4,4,4)	1	1/70	1/70	1.43
1	dhyper(1,4,4,4)	16	16/70	17/70	24.2
2	dhyper(2,4,4,4)	36	36/70	53/70	75.71
3	dhyper(3,4,4,4)	16	16/70	6 /70	.57
4	dhyper(4,4,4,4)	1	1/70	70/70	100.00

```
>
> # lotto
> # 6 right out of 49 (= "lotto gambling")
> teataste(49,6)
[1] 13983816
>
> # 1 of 13983816
> choose(8,4)
[1] 70
```

En el experimento hay diferentes posibilidades: entre cero y 4 correctas, es decir, 5 resultados posibles. Se puede calcular en R utilizando la distribución hipergeométrica (función de densidad) (véase la Tab. 4.1, véase el código R en `ptII_quan_classicstats_Fisher_ladyteataste.r`). La función devuelve el número de combinaciones. Los argumentos son

- los resultados correctos del experimento (es decir, la determinación correcta de "primero el té" o "primero la leche")
- el número de todos los "té primero"
- el número de todos los "primero la leche"
- el número de sorteos

Por ejemplo, para tres asignaciones correctas, la probabilidad sería p = 16/70, y para tres o cuatro asignaciones correctas, p = (16 + 1)/70 ≈ 0,2429.

En R, el experimento puede calcularse directamente a partir de una tabla de frecuencias utilizando fisher.test() (`ptII_quan_classicstats_Fisher_ladyteataste.r`).

```
> # use Fisher test to calculate prob
> # for '8 right out of 8' with fixed margins
> tea.test <- matrix(c(4, 0, 0, 4), nrow=2,
```

```
+ dimnames = list(Guess=c("Milk","Tea"),
+ Truth=c("Milk","Tea")))
> tea.test
Truth
Guess Milk Tea
Milk 4 0
Tea 0 4
> tea.ftest <- fisher.test(tea.test, alternative = "greater")
> tea.ftest
Fisher's Exact Test for Count Data
data: tea.test
p-value = 0.01429
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
2.003768 Inf
sample estimates:
odds ratio
Inf
> # check
> all.equal(tea.ftest$p.value,1/70)
[1] TRUE
> # = 1/70
```

¿Cómo resultó el experimento? Fisher no escribe nada al respecto, pero según Salsburg (2001), Lady Bristol fue sorprendentemente capaz de emparejar correctamente las 8 muestras de té.

4.3.3 Teoría de Neyman-Pearson - comportamiento inductivo, planificación deductiva

> „Is it more serious to convict an innocent man or to acquit a guilty? That will depend on the consequences of the error; is the punishment death or fine; what is the danger to the community of released criminals; what are the current ethical views on punishment? From the point of view of mathematical theory all that we can do is to show how the risk of errors can be controlled and minimized. The use of these statistical tools in any given case, in determining just how the balance should be struck, must be left to the investigator.“ (Neyman & Pearson 1933, S.296)
> ("¿Es más grave condenar a un inocente o absolver a un culpable? Eso dependerá de las consecuencias del error; si el castigo es la muerte o la multa; cuál es el peligro para la comunidad... de los delincuentes liberados; ¿cuáles son las opiniones éticas actuales sobre el castigo? Desde el punto de vista de la teoría matemática todo lo que podemos hacer es mostrar cómo se puede controlar el riesgo de errores y minimizada. El uso de estas herramientas estadísticas en un caso concreto, a la hora de determinar cómo debe alcanzarse el equilibrio, debe dejarse en manos del investigador." (Neyman & Pearson 1933, p.296)

Mientras que el enfoque de R.A. Fisher sigue la lógica de la cognición inductiva y la inferencia inductiva es generalmente criticada por Neyman (1955), la teoría de Neyman-Pearson (Willmes, 1996; Birnbaum, 1977) se centra en las decisiones inductivas "ciegas" en el contexto del razonamiento deductivo. No se trata de la validez, ni siquiera de la veracidad, de las hipótesis o intuiciones, sino de minimizar las entradas erróneas (tasas de error α o β, véase el cuadro 4.2 sobre el esquema de cuatro campos). Las decisiones se toman de forma rigurosamente "inductiva" a partir de los datos y en el marco de unas condiciones marco claramente especificadas (tamaño de la muestra, tasas de error α o β, potencia o fuerza de efecto).

Con Neyman-Pearson no hay hipótesis nula, sino dos hipótesis alternativas H_0 y H_1. La prueba decide qué consecuencias prácticas hay que sacar: por H_0 o por H_1. Este es el contexto en el que se origina el lema Neyman-Pearson (Neyman & Pearson, 1933, véase Capítulo 4.3.3.4). Así pues, esta lógica es excelente para la gestión o el control de la calidad, pero sólo es de ayuda limitada – si adoptamos el punto de vista de Fisher o Tukey – para generar conocimiento científico y descubrir algo nuevo. Acotar con precisión las condiciones de ensayo antes de la investigación permite tomar decisiones. Sin embargo, no fomenta el conocimiento que podría tener lugar fuera de estas condiciones marco y no se basa en la deducción (véase el capítulo 2.2). El diseño también requiere una definición clara de la muestra, a la que se hace referencia una y otra vez.

Tabla 4.2: Esquema de cuatro campos (tasas de error α-/Tipo I o β-/Tipo II)

		Realidad	
		H_0 verdadera	H_0 falsa
Testo	H_0 verdadera	OK	*a*-Tasa de error (falsos positivos)
	H_0 falsa	ß-Tasa de error (falsos negativos)	OK

En la teoría de Neyman-Pearson, no se trata sólo de encontrar algo raro desde el punto de vista estadístico, sino que el acontecimiento raro también debe ser significativo desde el punto de vista práctico. La significación práctica (potencia de efecto) y, posteriormente, la fuerza de la prueba, es decir, la capacidad de un procedimiento de medición para detectar diferencias, desempeñan un papel de apoyo en la teoría de Neyman-Pearson. Esto distingue claramente el enfoque de Fisher y añade una dimensión de significación práctica en lugar de "meramente" estadística que Fisher no tuvo en cuenta en absoluto. Es importante tener en cuenta que las entradas pueden anularse entre sí. Si la potencia de una prueba es baja, necesita un efecto mayor o una muestra más grande para poder tomar decisiones. La potencia o fuerza de una prueba estadística resulta de la relación entre la potencia y la tasa de error β y representa la probabilidad de decidir a favor de H_1 y en contra de H_0 si H_1 es válida. Por tanto, la potencia y la tasa de error β son diametralmente opuestas (véase el cuadro 4.2):

Poder de la prueba (Power) = 1 - β (4.3)

4.3.3.1 Significación práctica en el contexto de la significación estadística

Un acontecimiento "significativo" poco frecuente no tiene por qué tener una gran importancia en la realidad. Por ejemplo, un eclipse solar total en un punto concreto de la Tierra puede ser un acontecimiento muy poco frecuente, que sólo se produzca cada varios cientos de años o incluso menos. Sin embargo, si esto ocurre, poco cambia para toda la humanidad en la actualidad y los beneficios prácticos son extremadamente modestos, a menos que usted sea un ávido astrónomo o tal vez supersticioso. Por el contrario, sucesos muy comunes pueden ser extremadamente significativos (por ejemplo, la respiración): ocurren continuamente hasta que morimos y son vitales y, a menudo, no nos damos cuenta de la presencia de la respiración. Por ejemplo, inspiramos y espiramos alrededor de 500 millones de veces en una esperanza de vida de algo menos de 78 años, y a ritmos variables: los niños respiran más rápido, los adultos más despacio... y esto cambia en función del esfuerzo, el estado de salud, etc. Raro y frecuente son, pues, categorías que no se definen de forma absoluta, sino en relación con un acontecimiento concreto, en función de las expectativas, etc. Raras y frecuentes son, pues, categorías que no est-n absolutamente definidas, sino en relación con un acontecimiento concreto, en función de las expectativas, etc. Raras y frecuentes son, pues, categorías que no est-n absolutamente definidas, sino en relación con un acontecimiento concreto, en función de las expectativas, etc. ¿Sucede algo en poco tiempo o a lo largo de mucho tiempo? Y el sentido existe como en una dimensión independiente. Que algo sea significativo no tiene que ver necesariamente con el hecho de que ocurra rara vez o con frecuencia. Un acontecimiento singular sin mayor interés desde el punto de vista de la estadística clásica (como una sola frase en una conversación terapéutica) podría ser decisivo para la vida y ni siquiera tiene por qué ser repetible, es decir, replicable. Pero ciertamente los procesos implicados pueden reconstruirse cualitativamente de tal manera que las consecuencias del acontecimiento singular puedan entenderse a nivel de la vida práctica.

La importancia práctica se define por las consecuencias prácticas reales y no por la determinación de la rareza estadística. Es la medida de la diferencia o asociación, ya sea en la escala original o en un formato

estandarizado como la d de Cohen (Cohen, 1962, 1992), la r de correlación de Pearson o la η2 (Cohen, 1969; Richardson, 2011) en la computación de un anova.

Si dirigimos nuestra atención a la búsqueda de diferencias entre dos grupos, encontrar diferencias no es sorprendente, porque el mundo cambia constantemente y la mayoría -¿quizá incluso todos? - depende de complejas relaciones causa-efecto. Las diferencias parecen ser la norma, más que la excepción, cuanto más de cerca se observa un objeto. Con el aumento de la precisión, conceptos como el de igualdad se pierden y son sustituidos por la cuestión de la magnitud de las diferencias y su significado. Tukey (1991, p.100) comenta este tema:

> „All we know about the world teaches us that the e ects of A and B are always different — in some decimal place — for any A and B. Thus asking 'Are the effects different?' is foolish".

Por el contrario, se da el caso de que las diferencias prácticamente significativas se encuentran relativamente bien incluso con tamaños de muestra pequeños. Y luego la cuestión es si la dirección y la magnitud de las diferencias son coherentes con las expectativas postuladas, que Gelman y Carlin (2014) subrayan. Por el contrario, las diferencias prácticamente insignificantes pero numéricamente existentes requieren tamaños de muestra muy grandes para que la rareza estadística sea detectable.

Esperar constancia o no variabilidad aquí parece casi inútil o incluso extremadamente raro. Cuanto más de cerca miremos (por ejemplo, con un láser), más claramente se distinguirán unas estructuras de otras. El mundo se ve muy distinto a través de un microscopio que a través de nuestros ojos, porque la resolución es diferente. Sin embargo, sigue siendo el mismo mundo, pero puede que con el aumento de la precisión se pase por alto la "visión de conjunto". Un aumento de la precisión para encontrar diferencias -utilizado aquí como sinónimo de significación estadística- no significa necesariamente un aumento de la cognición significativa, es decir, que las diferencias encontradas representen información significativa. El conocimiento significativo se basa en mucha información e incluye una buena pregunta, teoría, diseño, implementación y, por último, pero no por ello menos importante, una aplicación en la realidad. La figura 4.1 muestra la correlación entre la potencia y la fuerza del efecto E (`ptII_quan_classicstats_-N-P_powerfunc.r`).

```
# power function - different effect sizes
alpha <- 0.05
N <- 100
# power <- 0.8
ES <- seq(-1,1,0.01)
# calculate power based on other values (see above)
res1 <- pwr.t.test(n=N,d=ES,power=NULL,sig.level=alpha,
type="two.sample",alternative="less")
res2    <-  pwr.t.test(n=N,d=ES,power=NULL,sig.level=alpha,
type="two.sample",alternative="greater")
res3    <-  pwr.t.test(n=N,d=ES,power=NULL,sig.level=alpha,
type="two.sample",alternative="two.sided")
```

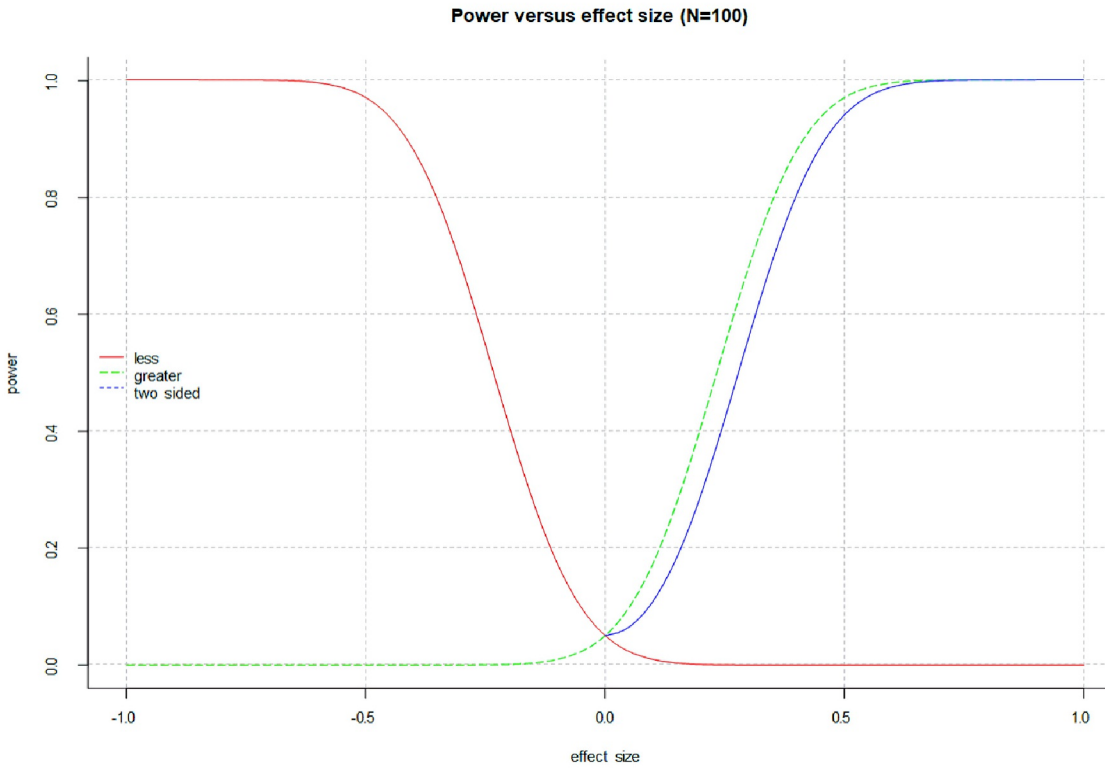


Figura 4.1. *Potencia frente al fuerza del efecto*

```
# plot power functions vs. effect sizes for less/ greater/ two-sided
kol <- rainbow(3)
plot(res1[["d"]],res1[["power"]], panel.first=grid(), bty="l",
       type="l",lty=1,col=kol[1],
       main=paste("Power versus effect size (N=",N,")",sep=""),
       xlab="effect size",ylab="power")
lines(res2[["d"]],res2[["power"]],type="l",lty=2,col=kol[2])
lines(res3[["d"]],res3[["power"]],type="l",lty=3,col=kol[3])
legend("left", col=kol,lty=c(1,2,3),
       legend=c("less","greater","two sided"), bty="n", cex=0.9)
```

Un pequeño ejemplo ilustra la diferencia entre significación estadística y práctica. Tomamos dos muestras, de las que determinamos de antemano los parámetros teóricos de la población. Como antecedente contextual elegimos el intervalo CI, que viene dado usualmente por $\mu = 100$ y $\sigma = 10$. La primera muestra tiene un CI medio "verdadero" de 100 puntos, la segunda de 104 puntos. La *d* de Cohen entre las poblaciones es $\delta = 0{,}4$. La diferencia absoluta de las dos distribuciones es de 4 puntos de CI. La desviación típica σ es idéntica en ambas poblaciones con $\sigma = 10$. Primero generamos una muestra aleatoria de ambas distribuciones con $n = 30$ personas cada una y las comparamos con una prueba *t* de dos muestras de forma clásica-estadística (`ptII_quan_classicstats_N-P_stat-signif-isNOT-practsignif.r`):

```
> # define population parameters
> mu1 <- 100
> mu2 <- 104
> sigma <- 10
> # seed for random numbers to replicate results > seed <- 9876
> set.seed(seed)
> # first sample
>n1<-30
> samp1 <- rnorm(n=n1, mean=mu1, sd=sigma)
```

```
>   samp2 <- rnorm(n=n1, mean=mu2, sd=sigma)
>   summary(samp1)
Min. 1st Qu. Median    Mean 3rd Qu.   Max
88.78 96.83 104.50 103.16 108.58 118.10
> # check for identical d's based on different calcs of pooled sd
> cohensd(samp1,samp2, check=TRUE)
Fehler in cohensd(samp1, samp2, check = TRUE) :
unbenutztes Argument (check = TRUE)
> t.test(samp1, samp2)
Welch Two Sample t-test
data: samp1 and samp2
t = -0.48271, df = 49.544, p-value = 0.6314
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-6.735432 4.125799
sample estimates:
mean of x  mean of y
101.8589   103.1637
```

Obviamente, las dos distribuciones no difieren en la α habitual = 0,05 de error entre sí (p = 0,63). Los valores determinados empíricamente de las distribuciones aleatorias ya se aproximan bastante a los valores de la población ($\bar{x}1$ = 101,86 para un grupo y $\bar{x}2$ = 103,16 para el otro). El valor absoluto de la d de Cohen es d = 0,12. Si repetimos lo mismo con los mismos números aleatorios pero una muestra el doble de grande, el valor p de p = 0,02 cae por debajo del llamado exceso de probabilidad crítica de un nivel de error convencional del 5 %. Los demás valores, así como la d de Cohen, se aproximan lentamente – ¡y como era de esperar! – a los valores de la población.

```
> # second sample, same parameters, n2 = 2*n1 >n2<-60
> samp3 <- rnorm(n=n2, mean=mu1, sd=sigma)
>   samp4 <- rnorm(n=n2, mean=mu2, sd=sigma)
>   summary(samp3)
Min. 1st Qu. Median  Mean 3rd Qu.   Max
76.30 92.71 99.07 100.06 105.36 124.24
>    summary (samp4)
Min. 1st Qu. Median  Mean 3rd Qu.   Max
86.29 96.85 103.50 104.71 113.44 129.70
> cohensd(samp3,samp4)
d|mean sd d|pooled sd 0.4336792 0.4336792
> t.test(samp3, samp4)
Welch Two Sample t-test
data: samp3 and samp4
t = -2.3754, df = 117.36, p-value = 0.01915
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-8.5285586 -0.7733632
sample estimates:
mean of x mean of y
100.0632 104.7142
> #Cohen's d population
> cohensd.pop <- (mu2-mu1) / sigma
> cohensd.pop
```

No hemos cambiado nada en los cálculos, salvo duplicar la muestra. La diferencia real entre grupos se mantuvo constante y deberíamos reflexionar sobre ello. La figura 4.2 muestra ambas distribuciones, así como los parámetros poblacionales y los valores empíricos en cada caso.

```
#call
par(mfrow=c(2,1), mar=c(5,6,5,5), oma=c(2,1,1,1), "cex.axis"=0.8)
# plot 1
breite1 <- plot.lines(samp=samp1, pop.mw=mu1, mu2=mu2,
           sigma=sigma, d=cohensd.pop)
plot.lines(samp2, colo="blue", pop.mw=mu2, add=TRUE)
mtext(expression(paste("sample 1",sep="")), 4, line=3, cex=0.9)
```

```
# plot 2
plot.lines(samp3, breite=breite1, pop.mw=mu1, mu2=mu2,
           sigma=sigma, d=cohensd.pop)
plot.lines(samp4, col="blue", pop.mw=mu2, add=TRUE)
mtext(expression(paste("sample 2",sep="")), 4, line=3, cex=0.9)
# title
mtext(expression(paste("Comparison of two normal distributed samples",
           sep="")), 3, line=-2, cex=1.5, outer=TRUE)
mtext(expression(paste("with constants ",mu[1],", ",mu[2],", ",sigma,
           " and ",delta)), 3, line=2, cex=1.2, outer=FALSE)
# legend
par(fig=c(0, 1, 0, 1), oma=c(0, 0, 0, 0), mar=c(0, 0, 0, 0), new=TRUE)
plot(0, 0, type="n", bty="n", xaxt="n", yaxt="n")
# add a nice legend with information
legend("bottom", legend=c(expression(paste(mu," (= population)")),
           expression(paste(bar(x)," (= sample)")),
           expression(paste("max. density (= sample)")) ),
           xpd=TRUE, horiz=TRUE, inset=c(0,0),
           y.intersp=2.4,
           col="black", text.col="black", lty=1:3, lwd=c(2,2,2), bty="n", cex=1)
```

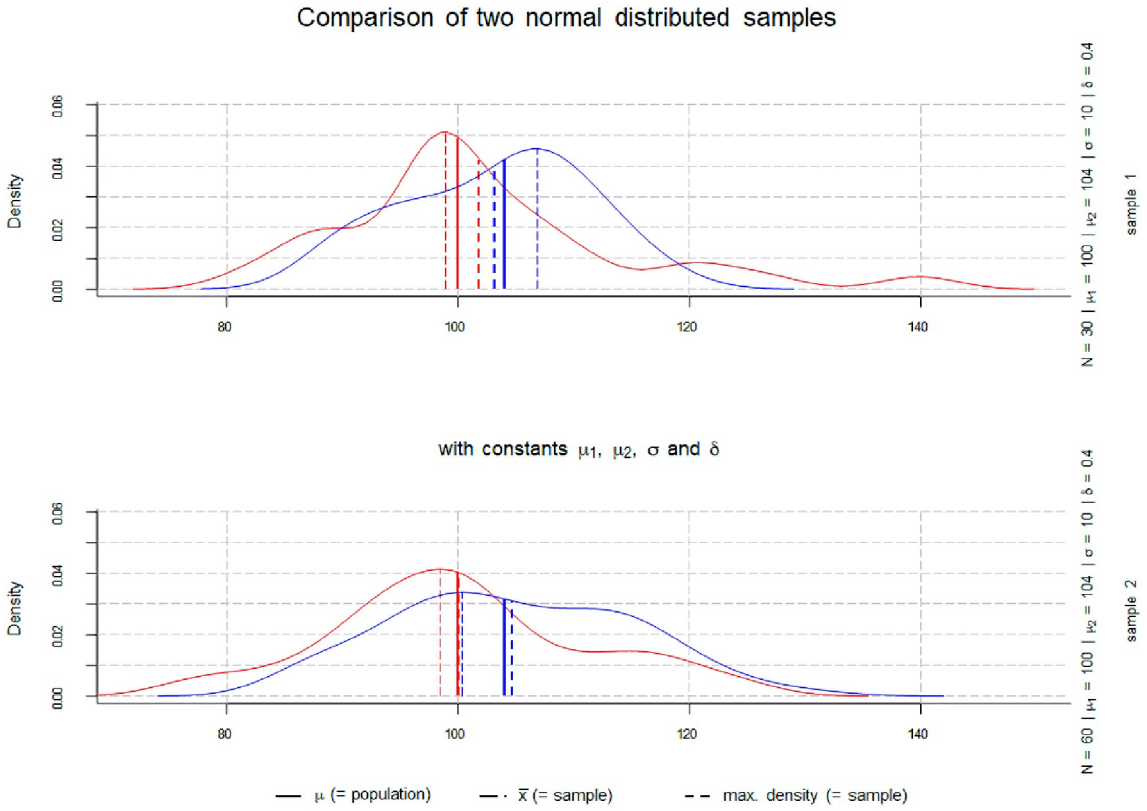


Figura 4.2 *Comparación de dos muestras con distribución normal*

Obviamente (véase Fig. 4.37, p.185) el valor p es una función directa del tamaño de la muestra. En cuanto al contenido, en principio no es muy esclarecedor. Según Cohen (1969), una $d = 0{,}4$ corresponde a un efecto pequeño o medio en el contexto de la prueba *t* (en R, véase cohen.ES() en el paquete pwr). Sin embargo, la *d* de Cohen es una cantidad abstracta sin una referencia concreta a la realidad. Si cambiamos a la escala original, una diferencia de CI de 4 puntos probablemente siga estando dentro del rango de variaciones comunes en la fiabilidad del retest (Deary, Pattie & Starr, 2013; Deary, Whalley, Lemmon, Crawford & Starr, 2000). Deary, Whiteman, Starr, Whalley & Fox (2004), por ejemplo, hallaron una estabilidad a largo plazo de las puntuaciones de rendimiento cognitivo entre los 11 y los 80 años de hasta

r = 0,69. A corto plazo, por tanto, cabe esperar correlaciones significativamente mayores con hasta $r > 0,9$. Todo ello sugiere que el CI no cambiará en órdenes mayores a corto plazo. Sin embargo, de ello no puede deducirse que sea completamente constante. Por tanto, los pequeños cambios se incluyen en la variabilidad esperada y 4 puntos de CI podrían formar parte de ella. Dado que en este ejemplo se conocen los valores de la población, sí que hay una diferencia de grupo que encontrar y si elegimos que el tamaño de la muestra sea infinitamente grande, siempre se encontrará. Pero todo esto sigue sin tener relevancia práctica para hacer afirmaciones razonables sobre lo que significa ahora en la realidad una diferencia de 4 puntos en un test de CI entre dos grupos, aunque como se ha demostrado esta diferencia sea "altamente significativa". ¿Significa esto ya que tendríamos que – por ejemplo en un entorno pedagógico – adaptar la enseñanza en función de los grupos? ¿Estamos ya ante una diferencia relevante en el potencial de logro? ¿Notamos siquiera una diferencia de 4 puntos de CI en la vida cotidiana? Podría ser, por ejemplo, que la diferencia encontrada indicara simplemente que el test está mal construido y, por tanto, favorece sistemáticamente a un grupo, que supuestamente tiene una puntuación de CI más alta. En otra prueba, también podría ocurrir lo contrario. Como vemos, sin una teoría razonable y bien fundamentada y sus correspondientes líneas de argumentación, no llegamos muy lejos.

Si realizamos un análisis de potencia a partir de los valores poblacionales, vemos que el estudio 1 con n_1 = 30 tiene una potencia = 0,33 y el estudio 2 con n_2 = 60 tiene una potencia = 0,58. Esto no nos impresiona mucho. Ahora no nos impresiona especialmente. Si preguntamos por el tamaño de muestra necesario en el marco de la teoría de Neyman-Pearson para encontrar el efecto descrito con una potencia de 0,90, necesitamos un tamaño de muestra de N = 132,31. Por tanto, se necesitan al menos 133 personas para encontrar de forma fiable una diferencia de grupo de 4 puntos de CI en el marco de los valores de población especificados y el exceso de probabilidad crítica α = 0,05.

```
> # power calculations
> crit.alpha <- 0.05
> #s ample 1 vs 2
> power.t.test(n=n1, delta=mu2-mu1, sd=sigma, sig.level=crit.alpha,
+          power=NULL, type="two.sample")
Two-sample t test power calculation
             n = 30
         delta = 4
            sd = 10
     sig.level = 0.05
         power = 0.3312747
   alternative = two.sided
NOTE: n is number in *each* group
> # sample 3 vs 4
> power.t.test(n=n2, delta=mu2-mu1, sd=sigma, sig.level=crit.alpha,
+          power=NULL, type="two.sample")
Two-sample t test power calculation
              n = 60
          delta = 4
             sd = 10
       sig.level = 0.05
          power = 0.5843645
     alternative = two.sided
NOTE: n is number in *each* group
> # N necessary
> power.t.test(n=NULL, delta=mu2-mu1, sd=sigma, sig.level=crit.alpha,
+          power=0.90, type="two.sample")
Two-sample t   test power calculation
              n = 132.3106
          delta = 4
             sd = 10
       sig.level = 0.05
          power = 0.9
     alternative = two.sided
NOTE: n is number in *each* group
>#N
> pwr.t.test(d=cohensd.pop, n=NULL, sig.level=crit.alpha,
+          type="two.sample", power=0.90, alternative="two.sided")
```

```
Two-sample t test power calculation
           n = 132.3105
           d = 0.4
   sig.level = 0.05
       power = 0.9
 alternative = two.sided
NOTE: n is number in *each* group
```

Si comparamos esto con los valores aleatorios generados, el resultado "significante" del experimento 2 parece aún más problemático, ya que posiblemente exista un producto aleatorio dada una potencia de ≈ 0,6. Si no intentáramos replicar esto y saber cómo justificarlo teóricamente de forma razonable, muy probablemente sobreestimaríamos el efecto o clasificaríamos mal la naturaleza del efecto. Deberíamos aclarar estas cuestiones teóricamente de antemano y los datos empíricos deberían respaldar estas afirmaciones iniciales. Con Gelman y Carlin (2014), podemos preguntarnos inmediatamente si la diferencia de grupo apunta en la dirección correcta y tiene la magnitud adecuada, especialmente cuando la potencia es baja y hay un resultado significativo. Deberíamos tener estas preguntas (véase el capítulo 4.3.3.2) preparadas por defecto para no ser presa de artefactos aleatorios. Por tanto, no sólo en la estadística cl-sica debemos preguntarnos inmediatamente cuando aparece la palabra "significance": "¿Esto es significativo en la realidad?" – en la escala original en la que se recogieron los datos; y "¿Cuáles son las consecuencias prácticas?", así como "¿Son la dirección y el nivel coherentes con las expectativas teóricas?". Tales preguntas deben formularse con una referencia sustantiva, lo que no impide resumirlas a un meta-nivel con fines de comparación con otros estudios, como se hace con las medidas comunes estandarizadas del tamaño del efecto. Utilizando la definición de Bateson (1985) del término informacional, podemos diferenciar aún más la cuestión.

> Tarea 4.2: Información
>
> Si la información es una diferencia que marca la diferencia, ¿qué diferencia hay aquí concretamente y cuál es su efecto?

La cuestión no es sólo si existe una diferencia desde el punto de vista estadístico, sino qué forma y tamaño adopta, qué calidad tiene y en qué dirección apunta, cuáles son las consecuencias de carácter práctico y si la teoría representa adecuadamente los datos y el modelo estadístico.

En general, *partimos de la base de que los modelos y las teorías son siempre erróneos en términos absolutos*. Son un reflejo de la realidad, forman parte de ella y no pueden situarse por encima de ella. Lo que nos interesa es saber si son herramientas útiles para describir nuestra realidad. Esto incluye si se ajustan a los datos (explicación y predicción) y conducen a conclusiones significativas y a la deducción de acciones en la realidad (intervención). Así que una vez más nos preguntamos por los criterios de la ciencia (véase el capítulo 1.2, Groeben y Westmeyer, 1981).

4.3.3.2 Dirección y magnitud: dos tipos de error infravalorados

El estadístico estadounidense Andrew Gelman y sus colegas proponen añadir a los tipos de error comunes de Neyman-Pearson α (tipo I "falsos positivos") y β (tipo II "falsos negativos", Eid, Gollwitzer y Schmitt, 2010, véase la Tab. 4.2) otros dos tipos de error

- tipo S(ign) y
- tipo M(agnitud)

(Gelman & Tuerlinckx, 2000; Gelman & Carlin, 2014). El error de tipo S examina *la dirección* de un efecto, es decir, si el signo es correcto y coincide con la expectativa. El error de tipo M describe si un efecto E también tiene *la magnitud correcta*. De este modo, la estadística clásica se aleja de la comprobación de

hipótesis nulas para adentrarse en la investigación de variables de influencia direccional y, de forma más general, en la modelización en un entorno complejo. Gelman (2004) lo explica en su blog de la siguiente manera,

> „A Type S error is an error of sign [...] I think it's fair to say that classical 2-sided hypothesis testing ?ts this framework: for example, if our 95 % interval for theta is [.1, .3], or if we say that theta.hat = .2 and is statistically signi?cantly di erent from zero, then our scienti?c claim is that theta is positive, not simply that it's nonzero.
> [...]
> A Type M error is an error of magnitude. I make a Type M error by claiming with confidence that theta is small in magnitude when it is in fact large, or by claiming with confidence that theta is large in magnitude when it is in fact small. The well-known problem of publication bias could lead to systematic Type M errors, with large-magnitude ?ndings more likely to be reported."

Por consiguiente, la atención ya no se centra en la significación estadística, sino en el comportamiento de los parámetros en el marco de condiciones contextuales definidas (variables). Por consiguiente, los autores ya no hablan de un análisis de potencia en el sentido de la teoría de Neyman-Pearson (decisión), sino de un análisis global del diseño para examinar los parámetros, sus tamaños y orientaciones. En un *análisis de diseño* de este tipo, no sólo se incluye información estadística, sino también toda la información contextual disponible procedente de la bibliografía, estudios empíricos conocidos, consideraciones razonables, etc. Esto se corresponde bastante bien con la recopilación de *conocimientos previos* en la estadística de Bayes (véanse los Capítulos 6.12 o 6.14.6). Esto significa, en particular, no reducir los tamaños de los efectos a muestras empíricas (singulares), sino ampliar la visión mediante una investigación exhaustiva a todo el conocimiento accesible en un área temática. Esto nos lleva a los métodos mixtos: las consideraciones cualitativas entran en la estimación del efecto numérico (tamaños) en el marco de la estadística clásica.

Gelman & Carlin (2014) proporcionan una función de R `retrodesign()` en el apéndice de su artículo. Se basa en la variable aleatoria d_{rep} con un modelo de probabilidad correspondiente, que estima cómo se comporta el efecto *d* en un hipotético estudio de replicación en las condiciones exactas del estudio original. La entrada de la función es

- el supuesto verdadero tamaño de efecto D - basado en una amplia variedad de fuentes externas de información,
- el error típico del parámetro estimado, y
- la tasa de error de tipo I supuesta α.

Además, el efecto d empírico observado y el valor p resultante son relevantes para el debate sobre el diseño (ibíd., p.644, Figura 1). Las siguientes cantidades son calculadas por `retrodesign()`:

- Potencia: la probabilidad de que el drep de replicación (valor absoluto) sea mayor que el valor que marca la significación estadística, es decir, el exceso de probabilidad crítica α.
- Tasa de error de tipo S: probabilidad de que la estimación replicada del parámetro tenga un signo erróneo cuando es estadísticamente significativa y diferente de cero.
- Probabilidad de error de tipo M (ratio de exceso): la expectativa del valor absoluto de la estimación del parámetro dividido por el tamaño del efecto cuando es estadísticamente significativo y diferente de cero.

El siguiente ejemplo de cálculo de Gelman y Carlin (2014) ilustra el procedimiento y la importancia de los parámetros. Los autores examinan críticamente un artículo de Durante, Arsena y Griskevicius (2013), que después de todo fue publicado en la revista Psychological Science (véase también la discusión en Gelman, 2013a o puramente en términos de contenido, Echidne, 2012). El objeto del estudio era el comportamiento electoral de las mujeres en EE.UU. en función de la fase del ciclo menstrual en la que se encontraban. Los resultados se indican en el artículo original en puntos porcentuales (por ejemplo, como diferencias). Por ejemplo, en un estudio preelectoral de 2012 se halló una diferencia d = 17 puntos porcentuales en el comportamiento de voto. Gelman y Carlin (2014) dudan del tamaño de este efecto y asumen que la medición

está muy sujeta a error. En primer lugar, se trata de una medición del cambio entre los individuos estudiados, y no dentro de ellos. La fase del ciclo menstrual criterio se preguntó sobre la memoria subjetiva y no se encuestó con precisión. El comportamiento electoral se encuestó a través de la imaginación. Esto significa que falta la comprobación de la realidad, es decir, si los encuestados acudieron realmente a las urnas y a quién votaron después. Según Gelman y Carlin (2014), la muestra era demasiado pequeña. Como resultado, ya existía un alto nivel de incertidumbre debido únicamente a esta información poco clara. El valor p notificado fue de p = 0,035 (prueba a dos caras, Durante et al., 2013, p.7), lo que, suponiendo una distribución normal, condujo a un valor z de |d/s| = z = 2,1 y, en consecuencia, a un error estándar de la estimación de SE = d/z = 17/2,1 = 8,1 puntos porcentuales (ptII_quan_classicstats_GandC_type-S-M-error.r).

```
# true effect size of 2, standard error 8.1, alpha=0.05
tes <- 2
# d difference = empirical mean
d<-17
# empirical reported two-sided p-value = 0.35
# abs(qnorm(0.035/2))
# =
zvalue <- qnorm(1-0.035/2)
zvalue
# standard error of difference
# d/s = zvalue
SE <- d/zvalue
SE
```

Sobre esta base, se realizó el análisis de diseño descrito para comprender mejor la situación de la información y sus limitaciones. En aras de la simplicidad, los errores de medición y los sesgos de selección señalados se dejaron de lado (temporalmente) y los valores empíricos se trataron como *válidos a primera vista*, lo que por supuesto no era el caso. Para ello, se realizó un análisis detallado de la bibliografía y los estudios sobre el cambio en el comportamiento electoral. Basándose en esta información previa, los autores concluyen que un límite superior plausible del efecto d esperado es d = 2 puntos porcentuales. En el artículo original se indica d = 17 puntos porcentuales. Esto supone una diferencia de un factor de 17/2 = 8,5. Este valor d = 2 se tomó como el presunto efecto verdadero y se introdujo como entrada en retrodesign() junto con el error estándar calculado anteriormente. Modificamos la función y le añadimos una salida gráfica (véase la Fig. 4.3). Estas modificaciones se inspiran en los ejemplos de Gelman (2014d) y Etz (2015a) (ptII_quan_classicstats_GandC_type-S-M-error.r).

```
> # values based on literature recherche (tes) and empirical p-value
> typsm.res.ref <- retrodesign(tes=tes, se=SE, graph=TRUE)
 Min.  1st Qu. Median Mean 3rd Qu. Max.
-31.3 -3.7     1.7    1.9  7.3     34.2
estimate
  n      missing  distinct   Info   Mean   Gmd
  10000        0      10000      1   1.856   9.132
    .05     .10        .25    .50     .75     .90
-11.536  -8.476     -3.654  1.717   7.265  12.278
    .95
 15.341
lowest : -31.3 -26.1 -25.7 -25.2 -24.2
highest:  29.8  30.2  30.8  31.6  34.2
> typsm.res.ref
$power
[1] 0.05707746
$typeS
[1] 0.2386568
$exaggeration
[1] 9.505829
> #empirical values based on author's empirical results
> retrodesign(tes=17, se=SE)
$power
[1] 0.5590079
```

```
$typeS
[1] 4.235435e-05
$exaggeration
[1] 1.335729
```

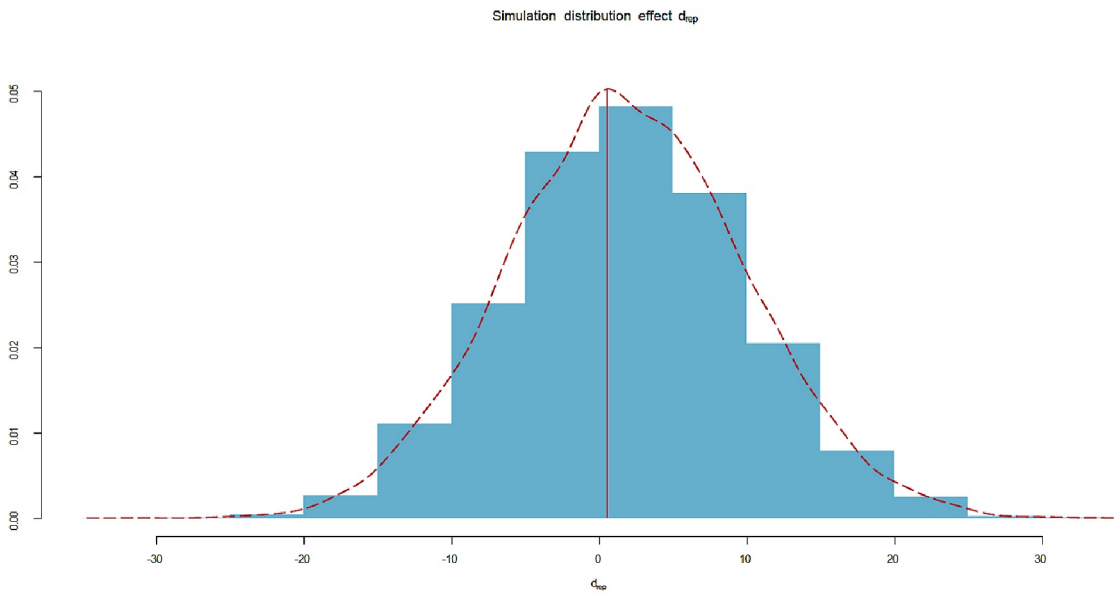


Figura 4.3 *Análisis de diseño – Durante et al. (2013, distribución de simulación del tamaño del efecto d_{rep}).*

Los resultados del análisis del diseño fueron:

- Potencia = 0.057
- Tasa de error de tipo S = 0.239
- Cociente de exceso tipo M = 9.506

Los resultados del análisis del diseño (véase la Fig. 4.3) indican que es muy probable que en un estudio de este tipo la estimación apunte en la dirección equivocada y – si es significativa – muestre una fuerte exageración de los patrones o relaciones que prevalecen realmente en la población. o relaciones realmente existentes en la población. Los autores subrayan que ese análisis del diseño no es sólo previo, sino también post-hoc a la realización de un estudio. En resumen, el resultado empírico estadísticamente significativo tiene poca sustancia para proporcionar información sobre las relaciones reales de interés de la población.

En este caso, resulta esencial realizar un trabajo previo de investigación del verdadero tamaño del efecto mediante búsqueda bibliográfica, meta-análisis de estudios relacionados, etc. Por un lado, esto supone un reto, pero por otro permite acotar exhaustivamente el tamaño del efecto desde diferentes perspectivas. Esta multi-perspectividad es prácticamente imposible debido a un estudio singular. La tasa de error de tipo S y la proporción de exceso de tipo M se producen especialmente cuando la media real de una población es pequeña y la variación de la muestra (error estándar) es muy grande. Si los estudios alcanzan la significación estadística en tal caso, suele haber tamaños del efecto sobreestimados. Estos también suelen tener el signo equivocado (véase también Etz, 2015a).

Por el contrario, esto significa que si existe una potencia elevada > 0.8 y prospectivamente un tamaño del efecto estimado de forma realista, se reducen drásticamente las dudas sobre un resultado significativo del estudio en relación con el signo del efecto y una posible sobre-estimación del valor empírico en relación con el parámetro poblacional. Gelman y Carlin (2014) subrayan, sin embargo, que precisamente estas condiciones – alta potencia, cálculo claro a priori del tamaño del efecto y análisis prospectivo de la potencia – *no* suelen darse en los estudios psicológicos. La situación es diferente para los estudios médicos, en los que una

potencia de 0:80 es un requisito común. Por el contrario, como puede verse en el estudio de Durante et al. (2013), los estudios psicológicos y de ciencias sociales se caracterizan por su baja potencia, que ya ha sido muy criticada en la literatura (Cohen, 1962; Sedlmeier & Gigerenzer, 1989) durante décadas. Además, hay muestras de pequeño tamaño y estimaciones del tamaño del efecto incorrectamente elevadas o incluso inexistentes antes de realizar un estudio. La figura 4.3 muestra un análisis de diseño de este tipo a lo largo del estudio descrito anteriormente. Contiene la distribución simulada de d_{rep} (histograma y estimación de la densidad, 10 000 simulaciones).

```
# retrodesign power plotting
D.range <- seq(0,50,0.1)
#call
typsm.res <- plot.power.retrodesign(typsm.res.ref=typsm.res.ref,
           D.range=D.range, tes=tes, se=SE)
```

El análisis de potencia (véase la Fig. 4.4) muestra a su vez en la curva superior que los problemas con el Tipo S de error comienzan a potencia < 0.1 y para el cociente de exageración de tipo M a potencia < 0.3 y se vuelven más drásticos a medida que disminuye la potencia.

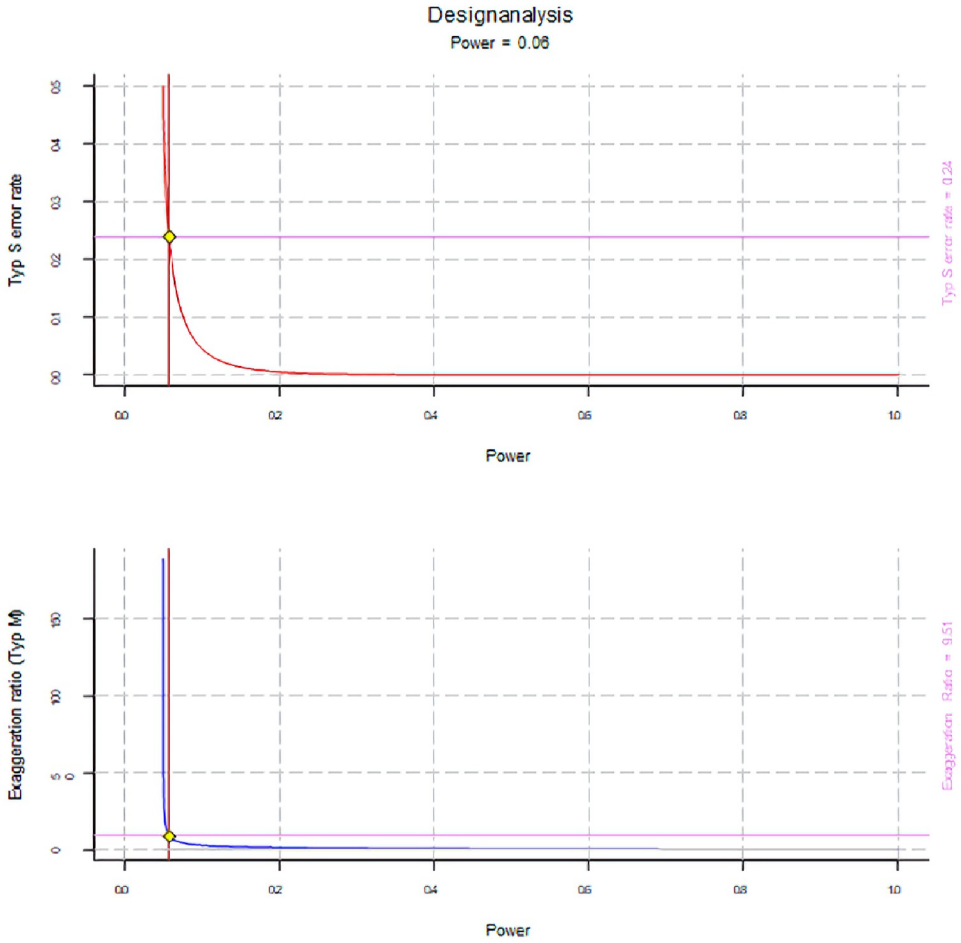


Figura 4.4 *Análisis del diseño de la simulación (potencia en función del tipo S o del tipo M).*

La línea horizontal violeta marca el valor empírico de la tasa de error de tipo S y el cociente de exageración de tipo M, respectivamente.

La línea vertical marrón marca la potencia basada en el análisis de diseño descrito anteriormente. La potencia máxima con una tasa de error de tipo S mínima de 0.1 viene dada por

```
> max(typsm.res[typsm.res[, "typeS"]>0.1, "power"])
[1] 0.07445166
```

Con un cociente mínimo de aumento de tipo de 2, una potencia máxima de

```
> max(typsm.res[typsm.res[, "exageración"]>2, "potencia"])
[1] 0.2480498
```

La figura 4.5 muestra la distribución del tamaño de los efectos según las distintas hipótesis. Ambos gráficos representan una distribución de muestreo en condiciones hipotéticas, es decir, cuando la distribución empírica de la muestra es la siguiente condiciones, es decir, si a partir de los valores del estudio empírico se se repetiría por simulación. La curva de la izquierda es la distribución del hipótesis nula $H_0 = 0$ y las áreas púrpura y naranja son los rangos de rechazo a un nivel de error común = 0.05. La línea vertical azul marca el la prueba contra cero, mientras que la línea roja marca el supuesto efecto verdadero d = 2 está marcado. El error estándar es SE = 8.1% puntos como se describe. Los puntos amarillos en 17% puntos denota la media de la muestra empírica. Bajo la hipótesis nula los rangos de rechazo se distribuyen simétricamente alrededor del punto cero.

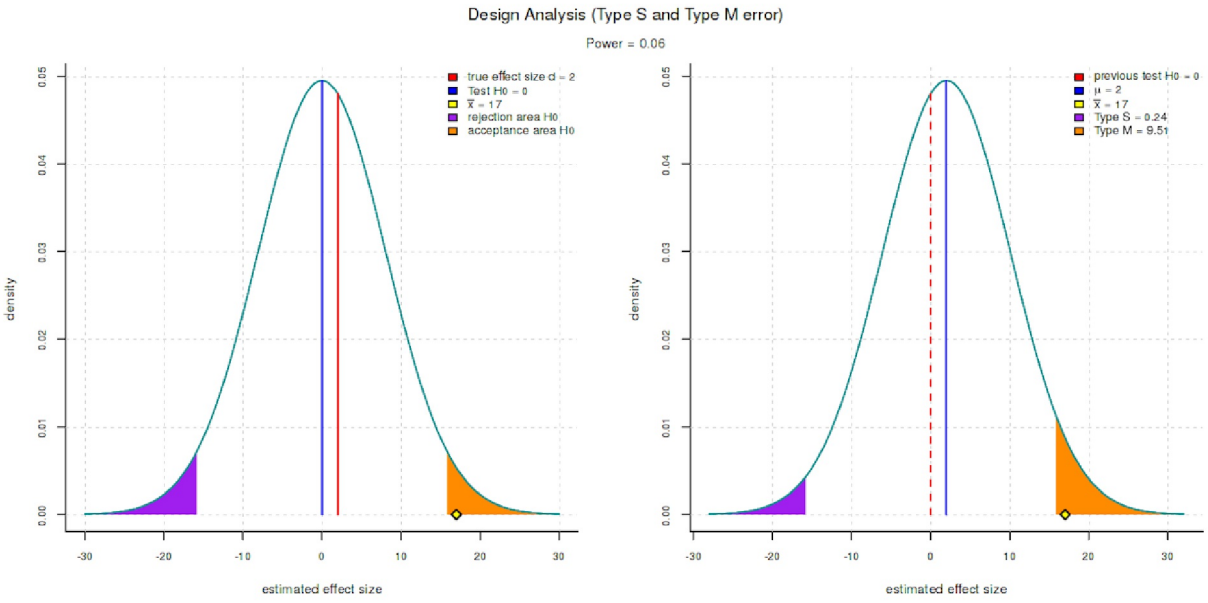


Figura 4.5. *Análisis del diseño (distribución de los tamaños de los efectos para $H_0 = 0$ bzw. = 2)*

```
# plot curves
plot.type.sm(typsm.res.ref=typsm.res.ref, range.dist=c(-30,30),
        mw=0, emp.mw=d, tes=tes, se=SE)
```

El gráfico de la derecha se basa de nuevo en las explicaciones de la verdadera media presunta de la población = 2 debido a la investigación y ya no en la prueba contra cero. Sólo con la desviación de cero podemos obtener un error de signo (tipo S) o de tamaño (tipo M).

Para ello, simplemente se desplazó la curva a la derecha en d = 2 puntos porcentuales sin cambiar nada más. Los rangos de rechazo permanecen constantes (en términos absolutos, es decir, los valores del eje X). Sin embargo, ahora tienen una densidad diferente (eje Y, frecuencia relativa). a ellos. Se ha perdido la simetría. Obviamente, la zona de la izquierda (morada) se ha reducido y el de la derecha (naranja) ha aumentado de tamaño. Dado que el error estándar con SE = 8.1 puntos porcentuales sigue siendo

significativamente mayor que la media poblacional supuesta = 2, cada resultado estadísticamente significativo debe tener un valor muestral significativamente mayor que el valor real de la población. Esto tiende entonces a acabar en el lado equivocado.

Sin embargo, las aplicaciones de los tipos de error S y M no acaban aquí. Gelman, Hill y Yajima (2012) señalan que el problema de las comparaciones múltiples causa problemas, pero en la práctica en las ciencias sociales los verdaderos efectos no son cero (véase también la paradoja de Mehl, sección 4.4.14.3). Una forma de abordar estos problemas es – además de las comparaciones planificadas y teóricamente justificadas – según Gelman, Hill y Yajima (ibid., p.209f.),

> „We prefer to frame the issue in terms of Type S or Type M errors. [...] we do not recommend classical methods that alter *p* values or (equivalently) make confidence intervals wider. Instead, we prefer multilevel modeling, which shifts point estimates and their corresponding intervals close to each other (that is, performs partial pooling) where necessary."

Así se obtienen estimaciones más fiables para los subgrupos.

4.3.3.3 *Modelos y respuestas: otros dos tipos de error*

Además de los errores de tipo S y de tipo M, la bibliografía pertinente menciona un error de tipo III y otro de tipo IV. Se refieren a la formulación del problema y, por tanto, a la operacionalización. La interacción entre la formulación del problema, el modelo empírico y los resultados (respuestas) se examina a partir del análisis de los datos. Es evidente que las mediciones o evaluaciones cuantitativas concretas de los errores respectivos son extremadamente difíciles o incluso imposibles. Así pues, se trata más bien de procesos de reflexión y directrices que deben ser a observar.

- El **error de tipo III** ("modelo equivocado, respuesta correcta") no se denota de manera uniforme en la literatura (para una visión general,Tate, 2015). Una versión común según Kimball (1957, p.134) lo denota como "el error cometido al dar la respuesta correcta al problema incorrecto" o, según Mosteller (1948, p.61), "rechazar correctamente la hipótesis nula para la respuesta incorrecta".

  La variante de Mosteller fue desarrollada por Kaiser (1960). Básicamente, se trata del error metodológico de haber formulado erróneamente el modelo básico -de modo que un nulo de modo que una confirmación estadística posterior no puede dar una respuesta significativa. respuesta. Esto ocurre cuando la operacionalización de la pregunta de investigación ya es defectuosa. Esto puede de las propias estadísticas. Como factores que influyen en los errores de tipo III se citan con frecuencia los siguientes mencionados con frecuencia:

  - trabajo teórico deficiente y/o explicaciones improvisadas poco convincentes de los resultados
  - operacionalización poco clara o incorrecta de los insumos y las variables
  - Identificación incorrecta de los factores causales en el modelo (es decir, no situar correcta-mente el foco en una investigación para poder generar una respuesta correcta y plausible a la pregunta de investigación; por ejemplo, confundir razones con efectos, etc.).

- El **error de tipo IV** ("modelo correcto, respuesta incorrecta") también existe en distintas variantes. Los de Marascuilo y Levin (1970) y Levin y Marascuilo (1972), siguiendo a Mosteller (1948), proponen denotar el error de tipo IV como la interpretación errónea de una hipótesis correctamente rechazada, lo que no significa otra cosa que extraer conclusiones equivocadas a partir de cálculos estadísticos correctos. Los factores significativos que influyen en el error de tipo IV son:

  - Utilización de procedimientos o pruebas de análisis de datos inadecuados para responder a hipótesis correctamente especificadas a partir de los datos.
  - Colineidades entre predictores y, por tanto, falta de claridad en la asignación de efectos y efectos causales. Efectos causales (se dan sobre todo en modelos de regresión múltiple)

- Falacia ecológica ("sesgo de agregación"): lo que es cierto para el grupo no lo es necesariamente para el individuo concreto. para el individuo concreto. Esto tiene que ver con la elección correcta de la unidad de investigación para para poder responder a una pregunta de investigación de forma adecuada al contexto.

Onwuegbuzie y Daniel (2003) enumeran los tipos de errores que pueden encontrarse no sólo en la investigación cuantitativa, sino también en la investigación cualitativa. Los autores enumeran los siguientes como errores típicos en el campo cuantitativo: la falta de comprobación de los supuestos estadísticos, la falta de discusión sobre la potencia y el tamaño de las muestras, tratamiento inadecuado de los datos multivariantes, el uso de procedimientos "paso a paso", la falta de valores de fiabilidad de las muestras anteriores o actuales y sin control de la tasa de error de tipo I. Sin embargo, lo que más llama la atención es la interpretación errónea de la significación estadística y los informes erróneos asociados y la mala interpretación de los intervalos de confianza y los tamaños del efecto.

Por el lado de la investigación cualitativa, no se legitiman adecuadamente los resultados de los estudios mediante la documentación de la fiabilidad y la validez, es decir, precisión y validez, además de abandono metodológico-analítico y filosófico. A esto se añade el problema de de generalización (véase también Gürtler & Huber, 2005), así como la incapacidad de evaluar cualitativamente el tamaño de los efectos e interpretarlos en consecuencia.

4.3.3.4 *Estudio de caso Neyman-Pearson – la gestión de calidad*

Tanto Fisher como Neyman-Pearson son representantes de la estadística frecuentista. La matemática subyacente no difiere entre las dos teorías, pero la interpretación y las variables utilizadas para las decisiones: H_0 y valor p para Fisher y H_0 frente a H_1, potencia, tamaño del efecto y tasas de error α y β para Neyman-Pearson. Un estudio de caso ficticio del ámbito de la gestión de la calidad demuestra ahora que funciona según Neyman-Pearson.

> Caso 4.1: Gestión de la calidad pt.1 – Piezas defectuosas.
>
> **Variante de caso 1**: Fabricamos piezas de plástico para la carrocería de un coche. Las piezas pueden ser defectuosas o no y se las venden o no. Esto da lugar a cuatro combinaciones (véase la tabla 4.3). Una pieza defectuosa es molesta, pero no pone en peligro la vida. En este sentido podemos como empresarios podemos hacer más hincapié en vender lo más posible. Las quejas son menos dramáticas, y en un cálculo de costes totales prevemos posibles reclamaciones y reservamos fondos para reparaciones o la sustitución de piezas defectuosas. Así, elegiríamos los niveles para las tasas de defectos α y β de tal forma que en la medida se vende lo posible y al mismo tiempo aceptar que las piezas defectuosas causarán ocasionalmente reclamaciones. Podemos reservar dinero de las ventas para ello. Con todo calidad, las ventas deberían ser significativamente superiores a los costes de reclamaciones y sustituciones. Entonces nuestro caso de negocio tendrá éxito. La atención se centraría claramente en vender las piezas potencialmente defectuosas y no en evitar la venta de piezas defectuosas. Ser un fabricante serio no significa descuidar la gestión de la calidad. Se trata de la estrategia económica básica, que está orientada a las grandes cantidades.

Así:

- Es bastante crítico *no* vender una pieza si se clasifica *erróneamente* como defectuosa (defecto de tipo I)
- Es menos dramático si se vende una pieza defectuosa (defecto de tipo II).

Esto representa una actitud empresarial común en el mercado de bienes de consumo. Basado en los dos hipótesis competidoras H_0 y H_1 resultan los dos porcentajes de error conocidos, los respectivos porcentajes

de error, las respectivas decisiones correctas y sus consecuencias, así como las implicaciones para la fuerza de ensayo (véase la tabla 4.3):

1. H_0 – la pieza no es defectuosa y se vende (tasa de error de tipo I).
   Falso rechazo de H_0: una pieza se clasifica como defectuosa, aunque en realidad no lo es. En consecuencia, no se vende. Esto debe evitarse a toda costa. Así ajustamos α y β de modo que son necesarias grandes desviaciones de la tolerancia para que una pieza se identifique como defectuosa por una buena razón.

2. H_1 - la pieza es defectuosa y no se vende, sino que se desecha o se reelabora (tasa de error de tipo II).
   Falso rechazo de H_1: una pieza se vende o instala como no defectuosa, aunque en realidad lo sea. En el peor de los casos, el cliente se queja de esta pieza defectuosa. Esto no es necesariamente bueno para la imagen de marca, pero no causa mucho daño a nadie, siempre que no ocurra todo el tiempo. Tal situación es legítima en los cálculos empresariales, lo que no significa descuidar la gestión de la calidad y la calidad del producto. Ajustamos más generosamente β (es decir, más grande) en comparación con α. Se trata de un procedimiento estándar, como señala Cohen (1992, p.156): "the [...] specification for power is .80 (so = 0.20), a convention proposed for general use."

3. Decisiones correctas
   En general, las decisiones correctas no plantean problemas. Utilizan la información disponible y conducen a conclusiones coherentes y justificadas, que son o deben ser conocidas en cada caso.

4. Implicaciones para la potencia de las pruebas
   Nuestro procedimiento de prueba debe tener un alto poder para distinguir con precisión las piezas defectuosas de las que no lo son. Este es el verdadero trabajo de la gestión de la calidad y, en realidad, siempre se aplica. Con un número muy elevado de piezas, podemos acercarnos relativamente a los verdaderos parámetros de la población – comparables a una simulación – de modo que los errores estándar se vuelven muy pequeños. Entonces las tasas de error de tipo I y de tipo II pueden ajustarse bien.

Tabla 4.3: Decisiones de ensayo en la gestión de la calidad (variante 1)

		Decisión de Ensayo	
		Parte defectuosa (Rechazo de H_0, pro H_1)	Parte no defectuosa (ningún rechazo de H_0, pro H_0)
Realidad	Parte defectuosa (H_0 falsa, H_1 verdadera)	OK	β (Error tipo II, falso negativo)
	Parte no defectuosa (H_0 verdadera, H_1 falsa)	α (Error tipo I, falso positivo)	OK

H_0 y H_1 se excluyen mutuamente. En función del resultado de la decisión sobre la prueba, el producto debe desecharse o puede venderse. Desde el punto de vista estadístico, Neyman-Pearson han establecido un lema (1933) que hace afirmaciones sobre la prueba óptima que decide entre dos hipótesis contrapuestas H_0 y H_1. Aquí, la prueba con potencia de prueba (power) óptima se identifica si se rechaza la hipótesis nula porque el cociente de verosimilitud (likelihood) f_0/f_1 cae por debajo de un valor determinado. f_0 y f_1 representan las densidades de probabilidad de las hipótesis respectivas H_0 y H_1. El *lema de Neyman-Pearson* relaciona la prueba de significancia estadística con la fuerza óptima de la prueba. La probabilidad crítica para una tasa de error de tipo I se fija en α, proporcionando así un límite superior para la tasa de error de tipo I. La prueba óptima minimiza entonces la tasa de error de tipo II β, ya que la potencia y la tasa de error β se excluyen mutuamente (véase la Ec. 4.3, p.78).

El valor *p exacto*, es decir, la base de la decisión de la prueba con respecto a la tasa de error de tipo I, *no* nos interesa en la toma de decisiones inductiva. El valor *p* sólo sirve de base para la decisión y no tiene más significado o aplicación aparte de eso. Se trata de saber si la probabilidad crítica debe ser inferior y si debe rechazarse H_0 (= aceptación de H_1) o no (= aceptación de H_0). La prueba se basa en α, mientras que

β resulta de la fuerza de la prueba y de la planificación preliminar de una investigación en el marco de un análisis de potencia a priori (Buchner, Erdfelder y Faul, 1996).

α no resulta directamente de los datos empíricos como β (mediada a través del valor *p*). Por lo tanto (Eid, Gollwitzer y Schmitt, 2010, cap. 8.4-8.6), la tasa del α-error puede controlarse directamente a través del umbral de significación crítica, pero no la tasa del β-error, ya que ésta incluye en principio todos los demás escenarios posibles (véase la Fig. 4.7). β depende del verdadero valor del parámetro investigado (distribución no central), que generalmente se desconoce en gran medida o incluso por completo cuando se asume H_1 (hipótesis alternativa) en contextos empíricos, particularmente en investigaciones singulares completamente desconocido. Además del tamaño del efecto, el tamaño de la muestra y la varianza de la característica examinada en la población tienen un efecto directo sobre el tamaño de la tasa de error β. Además, está el error de medición, en nuestro ejemplo la gestión de la calidad. En general, cuanto mayor sea la muestra, mayor será el efecto verdadero y el tamaño del efecto, y cuanto menor sea la varianza de la característica, menor será la tasa de error β.

Los porcentajes de error suelen compararse con las sentencias judiciales para aclarar las decisiones implicadas. Un falso positivo (α-tasa de error) conduce a la condena de un inocente, mientras que los falsos negativos inocente, mientras que los falsos negativos (β-tasa de error) supondrían absolver a un criminal. Ambos son legalmente y socialmente inaceptables, pero los porcentajes de error dependen entre sí. Si se pone el listón muy alto para que, en la medida de lo posible, no se condene a ningún inocente, al mismo tiempo es cada vez más probable que se absuelva a los culpables debido a la dificultad de las pruebas necesarias. Lo contrario también es cierto: si se condena a todo el mundo basándose en pruebas poco sólidas, muchos delincuentes serán condenados legalmente, pero también lo serán personas inocentes. Como la verdad rara vez se conoce en la práctica, el problema fundamental no es tan fácil de resolver, ni algorítmicamente ni de otro modo.

Sin embargo, en general, en los estudios empíricos no replicados existe un escenario incierto en cuanto a la tasa de error β, y es probable que esto sea así en casi todos los estudios educativos o psicológicos, con pocas excepciones. En el caso de series de pruebas largas, como en el ejemplo anterior de la gestión de la calidad, esto es diferente y la potencia puede calcularse relativamente bien sobre la base de grandes conjuntos de datos y, por tanto, todo el proceso puede ajustarse a los requisitos respectivos. La muestra se define con precisión en la gestión de la calidad, por lo que el marco de generalización es inequívoca. Una situación tan clara es también muy poco frecuente en los estudios pedagógicos y psicológicos – y si lo es, probablemente se trate más de experimentos de laboratorio y no estudios de campo. Las simulaciones sirven para determinar parámetros de antemano sobre la base de la información conocida. Con tal a priori análisis de potencia, se puede planificar y llevar a cabo una investigación con precisión.

La pregunta de si un análisis de potencia también es legítimo a posteriori suele responderse negativamente y calificarse de poco seria (Hoenig & Heisey, 2001). La razón es que un análisis post-hoc sugiere algo que ya está en los datos y, por tanto, en los resultados conocidos, y estimar la potencia a partir de datos empíricos implica inherentemente que es difusa, es decir, que está sujeta a errores de medición. Sin embargo, el objetivo de un análisis de potencia a priori es planificar una investigación con mucha antelación, cuando aún no se dispone de datos. Desde el punto de vista bayesiana, esto puede entenderse como un intento de incorporar a un estudio información preliminar que de otro modo no podría utilizarse en el curso de un análisis de datos debido a las limitaciones de la estadística clásica. Gelman (2019a, 2018) ofrece varias discusiones detalladas sobre el tema basadas en estudios concretos. Sin embargo, también existen argumentos contrarios a esta visión negativa unilateral de los análisis posteriores al diseño de datos (Gelman, 2017c). Gelman y Carlin (2014) hablan de análisis de diseño en lugar de análisis de poder para alejarse de los debates sobre la significación indecible. Brunner y Schimmack (2018a) describen aquí un procedimiento con código R (Brunner & Schimmack, 2018b) para separar la potencia media de la potencia "post-hoc" observada y hacer afirmaciones sobre la replicabilidad de los resultados del estudio. Además, las dependencias se discuten con el concepto de importancia. En consecuencia, se sigue una estrategia consistente en examinar las direcciones y los tamaños de los efectos (véase el capítulo 4.3.3.2, errores de tipo S y M) en lugar de si un efecto es estadísticamente significativo (Gelman & Hill, 2007). Desde este punto de vista, un análisis post-hoc del diseño puede ser muy útil, concretamente para explorar el contexto de una investigación y para considerar los posibles efectos de forma más realista en el contexto de las posibilidades. A este respecto, se hace una distinción: se rechaza el análisis de potencia post-hoc, pero no el análisis de diseño post-datos, con el fin de

aprender algo estadísticamente sobre los datos recogidos para utilizarlo con el fin de obtener información para futuras investigaciones y planificaciones. Del mismo modo argumentó Lenth (2007-07) y ofrece un código R para determinar la potencia post-hoc de las pruebas *t*, así como pruebas *F* de efectos fijos y aleatorios.

Resumamos – α y β no se puede derivar y calcular directamente de la otra variable, lo que sería elegante. Sin embargo, como ambos pertenecen a distribuciones diferentes y la distribución concreta a la que pertenece β suele ser desconocida ("distribución no central"), esto no es posible. Por lo tanto, el cálculo de β sólo es exacto si se conoce la distribución no central y, por lo tanto, se conoce los *verdaderos* parámetros de la población. Por este motivo, se realiza un análisis de potencia a priori para estimar estas variables desconocidas antes de disponer de los datos. Así pues, se puede planificar una investigación en este sentido según Neyman-Pearson.

Ahora podemos añadir un segundo escenario modificado de nuestro caso ficticio de gestión de la calidad y contrastarlo con la variante de caso 1:

> Caso 4.1: Gestión de la calidad pt.2 – Piezas no defectuosas.
>
> Imaginemos ahora que somos propietarios de una fábrica que produce piezas para la construcción de aviones. La producción de cada pieza es muy cara y ya hemos invertido en ella una gran cantidad de costes de desarrollo. Así que debemos vender lo que producimos. Pero cuando vendemos una pieza, es parte integrante de la seguridad de un avión, por ejemplo, el mando electromecánico del elevador. Así que debemos ser muy meticulosos antes de dar un parte. Una pieza defectuosa puede hacer que un avión se estrelle. Es dramático porque la gente puede morir. Así que se convierte en algo realmente caro, posiblemente perseguible; y por cierto, esto es extremadamente malo para la imagen de marca de uno y por lo tanto existencialmente relevante.

Si reconsideramos la primera hipótesis, en realidad tendríamos que intercambiar las hipótesis H_0 y H_1 por la segunda. Los objetivos se muestran contrarios a la variante de caso 1, en la que las grandes cantidades y un control de calidad existente pero inferior tuvieron un efecto decisivo. Para el caso de vender una pieza defectuosa (H_0, tasa de error de tipo I) es ahora más dramática que la de no vender una pieza no defectuosa (H_1, tasa de error de tipo II). Sin embargo, no vender una pieza no defectuosa sigue siendo bastante caro. En general, la tasa de error de tipo I se elige para que sea muy pequeña y la tasa de error de tipo II se elige para que sea cuatro veces mayor que la tasa de error de tipo I (Cohen, 1969). Por ejemplo, el NIH (= Instituto Nacional de Salud/ EE.UU.) espera lo siguiente cuando conceden fondos para la investigación: el instituto espera datos sobre una potencia de al menos el 80% en las aplicaciones, lo que corresponde a una tasa de β-error de $1 - 0.8 = 0.2 =$, es decir, un nivel de significación convencional $\beta/4 * 0.25 = \alpha = 0.05$. Gelman (2017b) critica esta dependencia implícita del significado y las expectativas poco realistas del poder. Según el autor, lo siguiente se aplica a los factores de influencia en escenarios reales de investigación:

- Los tamaños de los efectos suelen ser menores de lo esperado y de lo que se cree
- Los tamaños de los efectos en las publicaciones suelen estar generalmente sesgados (debido únicamente al sesgo de publicación, que favorece a los estudios significativos).
- Se añaden los errores sistemáticos (errores de medición, etc.)
- Las variaciones en las condiciones de investigación y los cambios a lo largo del tiempo hacen el resto (véase el estudio de Nosek, Spies y Motyl, 2012).

En definitiva, esto indica que en un contexto de investigación real no es nada fácil alcanzar realmente el 80% de potencia. Esto es frecuentemente mucho más bajo.

Todo esto ya no encaja en este nuevo escenario de gestión de la calidad, ya que podría poner en peligro la calidad del producto y, en consecuencia, la seguridad de los vuelos. Esto da lugar a la siguiente situación

de las tasas de error, de modo que α *indica la desviación del defecto* y β *la desviación de la normalidad*. El escenario normal (es decir, no defectuoso) sería la distribución no central, que por el momento se desconoce. En el fondo, se trata de una situación desfavorable que sólo podría resolverse con una gestión rigurosa de la calidad, aumentando la financiación de la investigación, recopilando muchos datos preliminares, etc. Al mismo tiempo, podríamos poner en marcha un procedimiento por etapas y mantener básicamente la lógica original de α y β (véase la Tab. 4.3, p.95).

1. H_0 – La pieza no es defectuosa y se vende (tasa de error de tipo I).

   Falso rechazo de H_0 – Una pieza se clasifica como defectuosa aunque no lo sea. Sin embargo, una pieza clasificada como defectuosa no debe salir de fábrica, ya que podría poner en peligro la aeronave. Esto es caro y, si ocurre más a menudo, la existencia de la empresa y no sólo su caso de negocio podría resentirse. α tendría que ajustarse de tal manera que la gestión de la calidad hiciera sonar la alarma no sólo en caso de desviaciones extremas con respecto al caso normal (= la pieza no es defectuosa), sino ya en caso de desviaciones muy pequeñas. A continuación se realizarían otros controles independientes, tras los cuales se autorizaría o no la salida de la pieza de la fábrica. Nosotros elegiría un valor α mucho mayor que el de los estudios científicos, por ejemplo, $\alpha = 0.2$ o incluso mayor, pero desde luego no según las convenciones habituales, como el 5 % o el 1 %. En este caso sería peligroso, ya que podría ocurrir que una pieza estuviera realmente defectuosa. Preferimos clasificar una pieza como defectuosa más rápidamente y someterla así a más pruebas por separado, que son más complejas, en lugar de simplemente venderla. En última instancia, el enfoque por etapas ayuda a aumentar la potencia, ya que entonces utilizamos, por ejemplo, procedimientos de medición más caros pero mucho más precisos para identificar un posible defecto.

2. H_1 – La pieza es defectuosa y no se venderá, sino que se desechará o reelaborará (tasa de error de tipo II).

   Rechazar falsamente H_1 – Una pieza defectuosa se clasifica incorrectamente como correcta y se instala. Si un avión se estrella como consecuencia de ello, los daños son inmensos. Aquí necesitamos una fuerza de prueba extremadamente alta para minimizar. Y necesitamos conocer la distribución no central. Con grandes cantidades de producción esto es posible como se ha mencionado. β podría entonces elegirse de forma que se aplicara una fuerza de prueba de Potencia = 99.9999% y $\beta = 0.0001\%$. Ahora tenemos que considerar cómo obtener tal fuerza de prueba – ciertamente no a través de la estadística, sino a través del ya introducido gestión escalonada múltiplemente de la calidad. Y de nuevo, no llegamos a ninguna parte con las convenciones. Si nos atenemos a la convención y elegimos $\beta = 4 * \alpha = 0.8$, tenemos una diferencia de 0.8/0.0001 – eso es un ¡factor de 8 000! – y el accidente de avión está preprogramado.

Las dos variantes de casos son dos ejemplos extremos ficticios y sólo pretenden ilustrar los procesos de reflexión y cuestionar hasta dónde podemos llegar con las convenciones. No es en absoluto trivial qué hipótesis y por tanto qué dirección de decisión de la prueba está asignado a α y β.

Así pues, nos enfrentamos al dilema de elegir las direcciones de las hipótesis H_0 y H_1, o bien elegir *tamaños completamente no convencionales* de las tasa de error o bien no conocer realmente la distribución no central y asumir así un riesgo empresarial. Con una experiencia cada vez mayor, conocemos las distribuciones y podemos reajustarlas. Pero el problema básico persiste. Si estamos en el nivel de una empresa incipiente (*Start-up*), por ejemplo, carecemos de experiencia y de datos empíricos en los que basar los cálculos. Ambos ejemplos muestran de forma impresionante que, a pesar de las duras críticas a la teoría de Neyman-Pearson o a la estadística clásica ortodoxa (por ejemplo, Jaynes, 2003; Tschirk, 2014; Kruschke, 2015b), se puede tener aplicaciones extremadamente útiles. Aparte de esto, todavía hoy hay defensores de este enfoque (Mayo y Spanos, 2006; Mayo, 2018).

Por supuesto, el ejemplo podría bayesianizarse, pero ¿por qué? Sobre todo en el caso de la producción continua, hay tal cantidad de datos que la gestión de la calidad puede ajustar muy bien las variables pertinentes (tasas de error de tipo I y de tipo II, intensidad de las pruebas, tamaño de la muestra, etc.). De hecho, esto equivale a aprender de la experiencia, una característica que normalmente no se concede a la estadística clásica. Por desgracia, en las ciencias sociales no se realiza un cálculo tan minucioso. Más bien se prescinde de un análisis de potencia a priori, la muestra sólo se denota vagamente y, por tanto, el alcance de los resultados se lleva al terreno de la vaguedad máxima; y las tasas de error se fijan qua convención sin

razón alguna. Además, ni siquiera no se mencionan en absoluto la tasa de error de tipo II y la fuerza de la prueba correspondiente. El resultado son estudios con una potencia muy baja (Sedlmeier & Gigerenzer, 1989) y, por tanto, con una significación muy limitada. Sólo formalmente mencionamos que la replicación – que sin duda implica grandes retos – no tiene lugar en la mayoría de los casos.

Como ya se ha mencionado, las publicaciones científicas recomiendan de forma generalizada que se utilice β aproximadamente el cuádruple de α (Cohen, 1969) – otra convención a tener en cuenta. Esta orientación aproximada debe decidirse caso por caso, como se ha descrito anteriormente. Los dos escenarios ficticios descritos muestran cómo los tamaños α y β pueden variar drásticamente en función de la finalidad de la aplicación. Ambos escenarios tienen consecuencias reales que no sólo afectan a las empresas. Por una vez, podemos concluir con un alto grado de probabilidad que la elección qua convención de las tasas de error es, en nuestra opinión, un completo disparate y carece de rigor científico si no se tiene en cuenta el problema concreto. Una vez más, establecer umbrales no es nada trivial. Considere cualquier problema de la vida cotidiana y determine los valores umbral a partir de los cuales un suceso pasa a ser significativo. Porque de eso se trata. ¿Cuál es el umbral crítico entre la insignificancia y la significación? Con un poco de estudio, se dará cuenta de que esto no existe, sino que hay un área crítica dentro de la cual el significado emerge lentamente. Muchas veces los límites son difusos.

4.3.3.5 *Contabilidad de costes totales*

Como buenos empresarios, nos vemos obligados a hacer una contabilidad de costes completa para determinar el argumento comercial. Para ello es necesario prever un posible siniestro y clasificar las variables que influyen. Las posibles estrategias en economía incluyen *minimax* (minimizar la mayor pérdida posible), *invarianza* (es decir, ciertas cualidades no deben cambiar en vista de la pérdida) o *minimizar* la pérdida media (minimizar el valor esperado de una función de pérdida definida). Son concebibles muchas otras variantes en función de los criterios requeridos. En nuestros ejemplos, se trataría de vender el mayor número posible de piezas o evitar estrictamente la venta de una pieza defectuosa. La idea de la contabilidad de costes totales da lugar a distintas variantes de decisión:

- Hipótesis H_0 y H_1
  - Dirección de la hipótesis: ¿cuál de las hipótesis H_0 o H_1 corresponde a cada caso?
  - Importancia en términos de contenido: ¿qué significan las tasas de error de Tipo I α y Tipo II β y cuáles son sus consecuencias prácticas? ¿Qué consecuencias empresariales prácticas se derivan de ello?
  - Conocimiento previo de α: ¿cómo se generan los datos (distribución) y existen conocimientos previos?
  - β incógnito: ¿qué sabemos de la distribución no central y cómo podemos obtener o generar datos al respecto?

- Tamaño de la muestra, es decir, cantidad de producción relativa a un periodo determinado (día, semana, ...) en el que se realiza la prueba. A partir de ahí, podemos generar una distribución básica a lo largo del tiempo, ya que los datos crecen continuamente. Al cabo de unos años o incluso de unos meses, en función del número de productos fabricados, disponemos de datos fiables. Es posible incluso prever las fluctuaciones causadas, por ejemplo, por variaciones de material de los proveedores u otras variables de entrada, y tomar a tiempo las precauciones correspondientes. Esto nos lleva a la gestión de la calidad y el control directo del proceso de producción.

- Fuerza del efecto, es decir, cuán grande es la diferencia entre H_0 y H_1 y, por tanto, entre las dos poblaciones sometidas a prueba, es decir, defectuosos frente a no defectuosos. Dependiendo del tipo de defecto, esto puede variar. Dependiendo de la calidad de nuestro proceso de producción, el tamaño de las muestras puede variar considerablemente. Todos estos datos son relevantes porque nuestra potencia de prueba debe activarse ante efectos muy pequeños, es decir, desviaciones del caso normal. En el caso de la producción de piezas de aviones, es esencial invertir en la potencia

de pruebas para distinguir las piezas defectuosas de las funcionales. La situación es muy distinta en la primera variante para la producción de piezas de carrocería, ya que en este caso las consecuencias son mucho menos dramáticas.

El resultado es un control riguroso:

- Implementación y reajuste del análisis de potencia a priori: nos referimos a esto como análisis de diseño en el sentido de Gelman y Carlin (2014) porque es un proceso continuo. Se amplía continuamente con datos para que siempre pueda ajustarse a los análisis y conclusiones de los datos, por ejemplo, si cambian las condiciones de producción, los criterios de calidad, etc. Este control de calidad comprueba el proceso de producción. En función de los resultados, pueden derivarse distintas intervenciones (por ejemplo, calendario, ritmo, cambio de proveedor, etc.).
- Cálculo de los intervalos de condensación de los parámetros medidos para garantizar una gestión continua de la calidad. Esto permite medir y predecir con precisión incluso pequeñas fluctuaciones a corto o largo plazo. Esto se aplica tanto a la calidad del producto como a otros aspectos de la producción (por ejemplo, formación de empleados y directivos) y las ventas/distribución.
- Anticipar el número de piezas vendidas, teniendo en cuenta las condiciones del mercado (ventas constantes, estacionales, etc.). ventas, estacional, etc.).
- Anticipación de las reclamaciones, que como se ha visto pueden variar drásticamente en función del tipo de producto. Deben constituirse reservas o contratarse seguros con este fin.
- ... y otras decisiones empresariales (relaciones públicas, publicidad, fidelización de clientes, imagen, ...), que aquí son menos importantes pero se basan sobre todo en la integración de información cuantitativa y cualitativa.

Hemos elegido deliberadamente estudios de casos del sector empresarial para la teoría Neyman-Pearson. En primer lugar, las consecuencias de una mala planificación son aquí mucho más reales que en el caso de las ciencias (sociales), donde las consecuencias de unos estudios empíricos deficientes son prácticamente nulas. Y el sector empresarial es, en nuestra opinión, exactamente el ámbito en el que la teoría Neyman-Pearson se encuentra en el lugar adecuado. La teoría Neyman-Pearson está totalmente basada en datos y, por tanto, se le ha atribuido con razón el atributo de *objetividad*, pero esto sólo es cierto hasta cierto punto en todos los ámbitos (véase la discusión, sección 4.3.7). Sin embargo, permite un ajuste directo de muchos parámetros en función de los demás y limitar así la incertidumbre con respecto a las decisiones futuras, como exigen los dos ejemplos de gestión de la calidad. Esta planificación previa en forma de análisis de potencia a priori funciona porque todas las variables seleccionadas son mutuamente dependientes y, en consecuencia, queda fijada la última variable libre. Esto sigue aproximadamente la lógica de *grados de libertad* (Eid, Gollwitzer & Schmitt, 2010, p.229f.). Una pregunta típica de un análisis de potencia a priori de este tipo es sobre el tamaño de la muestra.

El cálculo resultante, que se selecciona siguiendo estrictamente el procedimiento de análisis estadístico elegido posteriormente, proporciona un buen punto de partida para planificar una investigación. Para los dos ejemplos ficticios, podemos realizar el siguiente cálculo en R utilizando la prueba *t* de dos muestras (desviación de la normal). Elegimos una prueba *t* simple para las muestras dependientes por razones de simplificación, ya que cada una de ellas procede del mismo proceso de producción. En la práctica, resulta un procedimiento mucho más complejo, en el que un análisis de diseño es mucho más complejo que el siguiente código R ficticio simple y la visualización de las figuras 4.6 y 4.7, respectivamente. Independientemente de R, es posible realizar análisis de potencia a_priori con el programa GPower (Faul, Erdfelder, Lang & Buchner, 2007; Faul, Erdfelder, Buchner & Lang, 2009). El siguiente código R muestra algunos ejemplos de cálculos de potencia y las correspondientes comparaciones gráficas de NHST para la comparación de una muestra (véase la Fig. 4.6) o para la comparación de dos muestras mediante una prueba t (véase la Fig. 4.7) (`ptII_quan_classicstats_N-P_nulldist-hypotest.r`).

```
plot.H0(mu0=90, N=4, type="t", alternative="greater")
```

Tarea 4.3: Tamaño de la muestra

¿Qué tamaño debe tener la muestra – dado una potencia de efecto previsto *d* y niveles α y β – si una prueba tiene una potencia de *1-* β y se realiza tal o cual análisis estadístico?

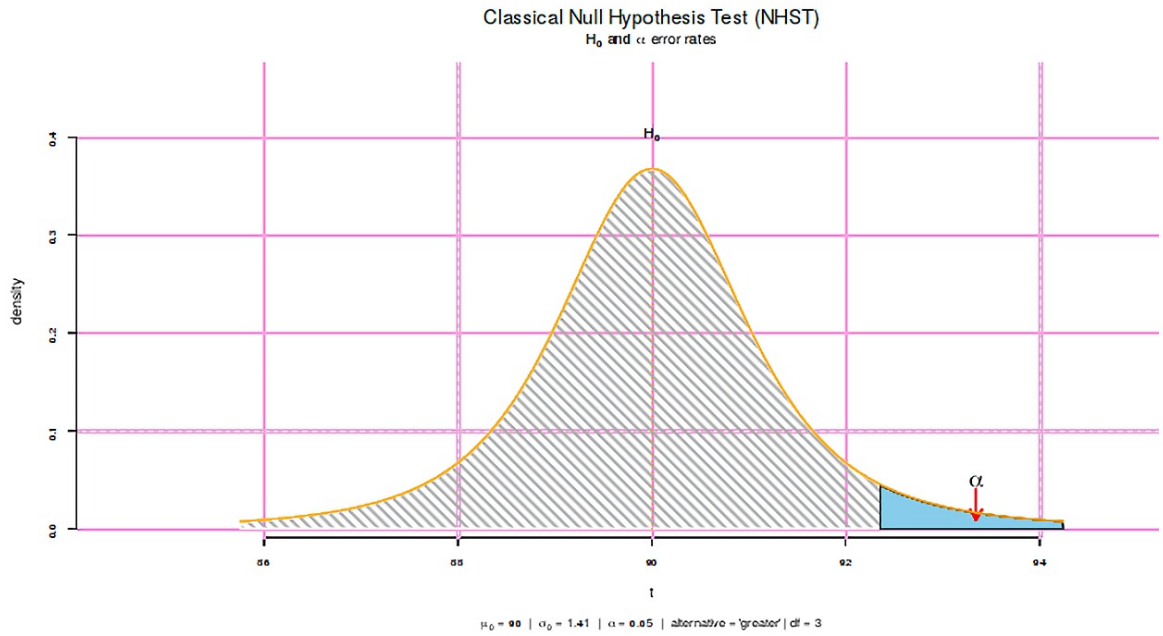


Figura 4.6: *NHST (una muestra, unilateral, más grande)*

Lo que funciona con la prueba de una muestra (base: distribución normal) también funciona con la prueba de dos muestras. En primer lugar, definimos las condiciones marco de las muestras:

```
# different calls
alpha.err <- 0.05
n1 <- 50
n2 <- 60
#mu1 <- 0
sigma1 <- 1
#mu2 <- 2
sigma2 <- 1
type <- "t"
#beta given, no mu1/ mu2/ delta
#delta = ?
beta.err <- 0.2
mu1 <- mu2 <- NA
delta <- NULL
alternative <- "two.sided"
```

y la llamada (base: distribución *t*):

```
plot.ab.err(n1=n1, n2=n2, mu1=mu1, sigma1=sigma1, sigma2=sigma2,
            delta=delta, alpha.err=alpha.err, beta.err=beta.err,
            type=type, alternative=alternative)
```

y el output:

```
Hypothesis testing with alpha and beta error-rate areas | type=' two.sided 'R-Output
n1 = 50
mu1 = 0
sigma1 = 1
n2 = 60
mu2 = 0.541
sigma2 = 1
df = 108
ES [d] = 0.541
lambda = 2.83
lambda [mean] = 2.85
lambda [sd] = 1.06
sd [pooled] = 1
alpha = 0.05
t [crit] = 1.98
beta = 0.2
Power [1-beta] = 0.8
type = t
df [comb.] = 104
alternative = two.sided
```

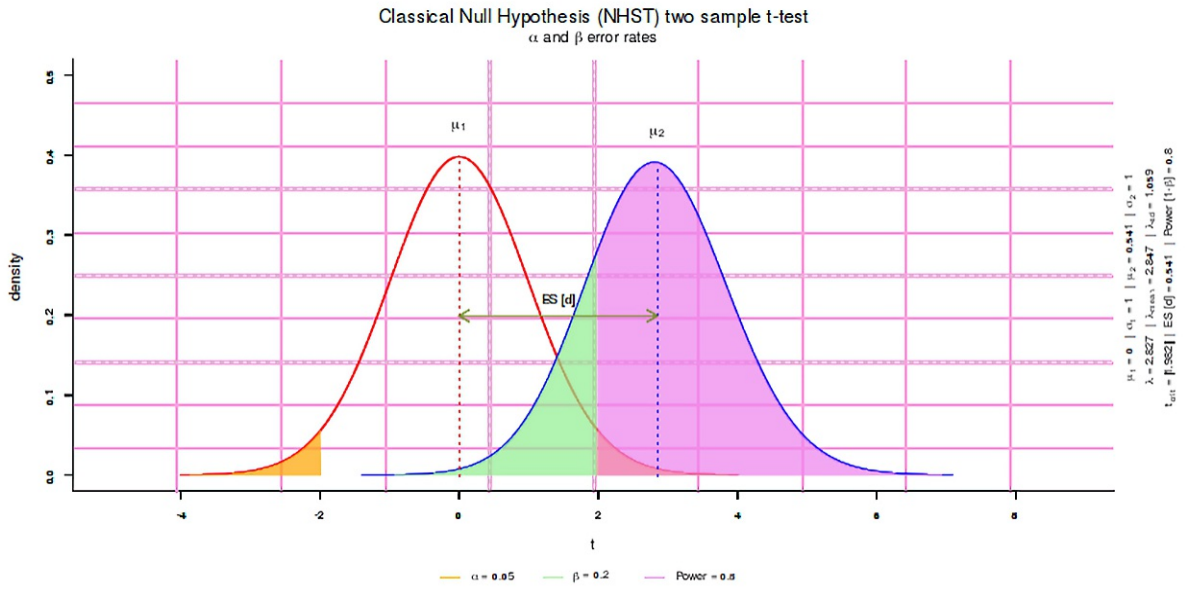


Figura 4.7: *NHST (dos muestras, bilateral, sin* μ*)*

Otras variantes son posibles resolviendo para el tamaño de la muestra u otros parámetros desconocidos. Tras los estudios de casos y la presentación básica del trabajo según Neyman-Pearson, pasamos a los subtemas específicos del enfoque. Empezamos con los *intervalos de confianza* y luego pasamos a la *relación entre estimación y prueba* (confianza frente a valor *p*) y *tamaño de la muestra*. Por último, examinamos la cuestión de la *subjetividad* en el contexto de la teoría Neyman-Pearson, ya que una mirada más atenta a los enfoques supuestamente definidos objetivamente demuestra que la objetividad no es lo que pretende ser. Más bien se incluyen muchas decisiones subjetivas, pero a menudo no se reconocen como tales (por ejemplo, las decisiones basadas en convenciones).

4.3.3.6 *Intervalos de confianza*

es decir, intervalos de confianza putativos según Neyman-Pearson (Neyman, 1935). Los *intervalos de confianza* son la contrapartida de los valores *p*. En la mayoría de los casos, aportan la misma información que una prueba de significación. Sólo que lo presentan de otra manera, dando la impresión de que contienen

información nueva. Este no es el caso, lo que no significa que hacen exactamente las mismas afirmaciones que los valores *p*. El intervalo de confianza de un parámetro estimado en la teoría de Neyman-Pearson no se corresponde con la comprensión intuitiva de la confianza. Uno suele entenderse como *la probabilidad de que un parámetro estimado (por ejemplo, la media, la diferencia de medias o el coeficiente de regresión) se encuentre en el intervalo de confianza especificado (por ejemplo, entre 80 y 100 puntos)*. Esta interpretación intuitiva corresponde a la definición bayesiana de intervalo creíble (Jaynes, 1976), pero no a un intervalo de confianza clásico según Neyman-Pearson. Oigamos hablar al propio Neyman (Neyman, 1937, p.349):

> „It will be noticed that in the above description the probability statements refer to the problems of estimation with which the statistician will be concerned in the future. In fact, I have repeatedly stated that the frequency of correct results will tend to α.* Consider now the case when a sample, E′, is already drawn and the calculations have given, say, θ (E′) = 1 and θ (E′) = 2. Can we say that in this particular case the probability of the true value of θ_1 falling between I and 2 is equal to α?
>
> The answer is obviously in the negative. The parameter θ_1 is an unknown constant and no probability statement concerning its value may be made, that is except for the hypothetical and trivial ones
>
> $$P\left\{1 \leq \theta_1^0 \leq 2\right\} = \left\{ \begin{array}{lll} 1 & \text{if} & 1 \leq \theta_1^0 \leq 2 \\ 0 & \text{if either} & \theta_1^0 < 1 \ \ or \ \ 2 < \theta_1^0, \end{array} \right\} \ \ldots \ (21)$$
>
> which we have decided not to consider."

Debe quedar claro que un intervalo de confianza específico no hace ninguna declaración de probabilidad sobre el *parámetro poblacional verdadero* relevante, por ejemplo, con qué probabilidad se encuentra o no dentro del intervalo de confianza, que a su vez se estima a partir de la muestra empírica, siempre que no se disponga de otra información. Esta cuestión probabilística relacionada con los parámetros puede responderse mediante *intervalos creíbles* bayesianos o IDH (= Intervalos de Alta Densidad) (Kruschke, 2011a, 2015b), que no existen en la teoría de Neyman-Pearson ni en la estadística clásica en general. Esperar tales afirmaciones de la teoría Neyman-Pearson la desacreditaría, ya que ni siquiera pretende hacer afirmaciones al respecto (Mayo, 1981). No obstante, la validez de los intervalos de confianza clásicos se discute de forma muy crítica (por ejemplo, Morey, Hoekstra, Rouder, Lee & Wagenmakers, 2016).

Así pues, existen varias interpretaciones erróneas de lo que se supone que son los intervalos de confianza frecuentistas. Los intervalos de confianza se refieren siempre a estadísticos, es decir, a parámetros de interés como medias, diferencias de medias, varianzas, etc. Aquí algunos ejemplos de lo que *no* son los intervalos de confianza. *No hacen declaraciones* sobre el hecho

- que con probabilidad *p* el intervalo de confianza estimado empíricamente contiene el verdadero parámetro poblacional.
- que el intervalo de confianza debe entenderse como una medida final de todos los valores plausibles del parámetro poblacional. Por otra parte, es legítimo interpretar los valores dentro del intervalo de confianza como valores plausibles para el parámetro poblacional, teniendo en cuenta el hecho de que las muestras se consideran aleatorias y, por tanto, sujetas a fluctuaciones.
- que el intervalo de confianza comprende el *p* % de los datos concretos de la muestra, es decir, que el *p* % de los datos de la muestra se encuentran dentro del intervalo de confianza especificado.
- que existe una probabilidad *p* de que, dado un intervalo de confianza *p* % estimado empíricamente, exista una probabilidad *p* de que el parámetro muestral de interés de una muestra futura esté contenido en este intervalo de confianza. Los intervalos de confianza hacen afirmaciones sobre el intervalo de confianza relativo al parámetro poblacional desconocido, no sobre los parámetros muestrales de un estudio relativos a los parámetros muestrales del siguiente estudio. Por otro lado, es legítimo afirmar que, con una probabilidad del *p* %, el intervalo de confianza estimado de un estudio futuro abarcará el verdadero parámetro poblacional. Aún no está claro de qué intervalos de confianza de qué estudios se tratará. Los valores de un estudio concreto se encuentran o no dentro del intervalo de confianza. No hay nada intermedio.

En cambio, las afirmaciones legítimas del intervalo de confianza son,

- que si un estudio se repite de forma idéntica en *p*% de los casos, los *intervalos de confianza* estimados a partir de las muestras – que, por cierto, varían de una muestra a otra – constituyen el verdadero parámetro poblacional.
- que si un estudio se repite de forma idéntica en el *100% = p% =* α% de los casos, los intervalos de confianza *no* incluyen el verdadero parámetro poblacional.
- que, para una encuesta futura, el intervalo de confianza estimado empíricamente incluya el parámetro de población con una probabilidad de al menos *p* en el intervalo entre los límites inferior y superior. Una vez recogida la muestra, esta probabilidad se concentra irrevocablemente en cero o uno.
- que el intervalo de confianza estimado empíricamente representa valores del parámetro poblacional de forma que no hay diferencia estadísticamente significativa con el parámetro empírico de la muestra al nivel *p%*. Sin embargo, esto debe utilizarse con cautela y criterio debido al fácil mal uso de los valores p y no es adecuado como única fuente de justificación de los efectos.
- que con el aumento del tamaño de la muestra el intervalo de confianza se hace cada vez más estrecho y converge asintóticamente – con el tamaño de la muestra hacía ∞ – al parámetro poblacional verdadero con una anchura de CERO y probabilidad UNO.

Queda totalmente abierto si en una investigación *concreta* el intervalo de confianza estimado a partir de la muestra incluye o no el verdadero parámetro poblacional. No hay ninguna probabilidad de que esto ocurra. O bien se encuentra dentro del intervalo, o bien fuera de él, y con la recogida de datos esto ya está determinado. No es posible hacer una declaración de probabilidad al respecto. Sólo es posible hacer afirmaciones basadas en la repetición de estudios idénticos (experimentos, exámenes, etc.); y todas las afirmaciones se refieren entonces a la totalidad de estas repeticiones. Las afirmaciones se refieren estrictamente a los intervalos de confianza, no a los verdaderos parámetros de la población. Esto convierte a la teoría de Neyman-Pearson en una teoría de la réplica, ya que deja claro que las afirmaciones se hacen sobre la base de réplicas de muestras y no sobre la base de estudios individuales. Las afirmaciones basadas en estudios individuales están sujetas a error porque son aleatorias.

Empíricamente, se deduce que para la determinación de los intervalos de confianza según Neyman-Pearson, por ejemplo, el diseño de un estudio se repite una y otra vez. Para cada estudio individual puede calcularse un intervalo de confianza – según la convención – de 95 % en torno a la respectiva variable empírica de interés (por ejemplo, la media, la pendiente de un parámetro de regresión o la diferencia de medias).

Lo siguiente se aplica a todos los intervalos de confianza del 95 % con la repetición de muestras correspondientes, que

- en el 95% de los casos (= *1* - α) se encuentran valores que encierran la verdadera diferencia de medias en la población, y
- en el 5% (= *tasa de error* α) de los casos, se encuentran valores para los que la verdadera diferencia de medias de la población está fuera del los intervalos de confianza.
- la anchura y el valor medio de cada intervalo de confianza determinado empíricamente están sujetos al azar en función de la muestra. Por lo tanto, varían como una variable aleatoria de una muestra a otra.
- para cada estudio, el intervalo de confianza calculado incluye o no la verdadera diferencia de medias poblacionales. No hay ninguna probabilidad de ello, como señala claramente Neyman en su artículo original (véase la cita anterior).

Un ejemplo ficticio: si se realizan repetidamente n = 50 estudios y se calcula un intervalo de confianza del 95% entre dos muestras (por ejemplo, diferencia de medias), en una media de 50 * .95 = 47.5 casos el intervalo de confianza encerrará la verdadera diferencia de medias λ; y en una media de 50 * (1 - .95) = 2.5 casos el intervalo de confianza *no encerrará* la verdadera diferencia de medias λ. Si a continuación calculamos el intervalo de confianza determinado empíricamente para uno solo de estos estudios, el intervalo

de confianza del 95 % en el ejemplo ficticio puede oscilar entre CI_{low} = 3.1 y CI_{up} = 3.5 puntos de diferencia de medias.

En general, un intervalo de confianza se calcula del siguiente modo, utilizando el ejemplo de la media aritmética:

$$p\% = 1 - \alpha \qquad \text{Relación entre CI y } \alpha \qquad (4.4)$$

$$CI_{\bar{x}} = \bar{x} \pm t_{\frac{\alpha}{2}} \cdot s_{\bar{x}} \qquad \text{Parámetro concreto (de la muestra)} \qquad (4.5)$$

$$CI_{\theta} = \theta \pm t_{\frac{\alpha}{2}} \cdot \sigma_{\theta} \qquad \text{en general} \qquad (4.6)$$

Un ejemplo de R lo demuestra (`ptII_quan_classicstats_N-P_confint.r`). Tomemos n_1 = 50, x_1 = 100 y s_1 = 10:

```
> # 95% N-P CI empirical mean
> n1 <- 50
> xbar1 <- 100
> sd1 <- 10
> ci.mean.res <- ci.mean(n1=n1, xbar1=xbar1, sd1=sd1, printshort=TRUE)
100 [97.16; 102.84]
> ci.mean.res
N  CI(low)  Mean CI(up)  CI(width) SD SE       CI(prob) t        df
50 97.15803 100  102.842 5.683937  10 1.414214 0.95     2.009575 49
> ci.mean(n1=n1, xbar1=xbar1, sd1=sd1, printshort=FALSE)
N  CI(low)  Mean CI(up)  CI(width) SD SE       CI(prob) t        df
50 97.15803 100  102.842 5.683937  10 1.414214 0.95     2.009575 49
> # rounded
> digits <- 2
> sapply(ci.mean(n1=n1, xbar1=xbar1, sd1=sd1), round, digits)
N     CI(low) Mean   CI(up) CI(width) SD
50.00 97.16   100.00 102.84 5.68      10.00
SE   CI(prob) t    df
1.41 0.95     2.01 49.00
```

A partir de los valores empíricos anteriores, los límites del intervalo de confianza teórico del 95% son CI_{low} = 97.16 y CI*up* = 102.84 con $t_{\alpha/2}$ = 2.01 (df = 49). Los intervalos de confianza pueden calcularse para cualquier cosa que pueda reproducirse según la teoría de Neyman-Pearson, por ejemplo, una simple diferencia de medias muestrales:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{(\bar{x}_1 - \bar{x}_2)}} \qquad (4.7)$$

con desviación estándar agrupada

$$t_{pooled} = \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}} \cdot \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}} \qquad (4.8)$$

con la desviación estándar de la diferencia de medias

$$s_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad (4.9)$$

y con desviación estándar agrupada

$$s_{pooled} = \frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2} \qquad (4.10)$$

y el intervalo de confianza

$$\mathrm{CI}_{(\bar{x}_1 - \bar{x}_2)} = (\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2}} \cdot s_{(\bar{x}_1 - \bar{x}_2)} \qquad (4.11)$$

y con varianzas iguales

$$\mathrm{df}_{(s_1=s_2)} = n_1 + n_2 - 2 \tag{4.12}$$

o con desiguales varianzas

$$\mathrm{df}_{(s_1\neq s_2)} = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{s_2}{n_2}\right)^2}{n_2-1}} \tag{4.13}$$

El cálculo de los grados de libertad para varianzas desiguales se basa en el intervalo t de Welch (Welch, 1947, 1951), en el que x_1, s_1, n_1, así como x_2, s_2, n_2 se determinan empíricamente y se estiman a partir de la muestra. La razón es que, por lo general, se desconocen los valores reales en cada caso porque no suele haber investigaciones preliminares exhaustivas. La probabilidad (es decir, la anchura) del intervalo sigue la convención habitual de 1 - α% = 95% o aprox. un valor t (cobertura de área unilateral) de 1.959964, redondeado a 1.96 e infinitos grados de libertad.

Tabla 4.4: *Intervalos de confianza - diferencia de medias (ejemplo ficticio)*

Muestra	N	$\bar{x}$	s
1	50	100.0	10.0
2	48	104.8	12.5

```
> # caclulate t-value
> prob <- .95
> alpha <- 1-prob
> qt(1-alpha/2, df=Inf, lower.tail=TRUE)
[1] 1.959964
```

Tomamos el ejemplo ficticio de una diferencia de medias entre dos grupos, suponiendo desviaciones típicas iguales y desiguales. La tabla 4.4 muestra los valores iniciales. En primer lugar, suponemos las mismas varianzas para ambas muestras y fijamos s_1 para ambas muestras:

```
> # 95% N-P CI empirical difference in means
> n2 <- 48
> xbar2 <- 104.8
> sd2 <- 12.5
> # sd1 = sd2
> ci.diff.in.means(n1=n1, xbar1=xbar1, sd1=sd1, n2=n2, xbar2=xbar2,
+           sd2=sd1, equal.var=TRUE)
             N  CI(low) theta CI(up) CI(width) SD SE
sample(1)   50  97.158 100.0 102.84 5.68      10 1.41
sample(2)   48 101.896 104.8 107.70 5.81      10 1.44
delta(mean) 98   0.789   4.8   8.81 8.02      NA 2.02
             CI(prob) t    df var(equal)
sample(1)   0.95     2.01 49 NA
sample(2)   0.95     2.01 47 NA
delta(mean) 0.95     1.98 96 TRUE
```

Ahora suponemos desviaciones estándares diferentes para ambas muestras (véase la tabla 4.4):

```
> # sd1 != sd2
> ci.diff.in.means(n1=n1, xbar1=xbar1, sd1=sd1, n2=n2, xbar2=xbar2,
+           sd2=sd2, equal.var=FALSE)
```

```
            N  CI(low) theta CI(up) CI(width) SD   SE
sample(1)   50  97.158 100.0 102.84 5.68      10.0 1.41
sample(2)   48 101.170 104.8 108.43 7.26      12.5 1.80
delta(mean) 98   0.245   4.8   9.35 9.11       NA  2.29
            CI(prob) t    df   var(equal)
sample(1)   0.95     2.01 49.0 NA
sample(2)   0.95     2.01 47.0 NA
delta(mean) 0.95     1.99 89.9 FALSE
```

Obviamente, los resultados difieren en esto, por lo que con diferentes varianzas los intervalos de confianza se amplían, como cabe esperar intuitivamente a medida que aumenta la incertidumbre. En el caso que nos ocupa, se trata al fin y al cabo de una diferencia con respecto a `CI(anchura)` para las dos `delta(media)` de

```
> 9.109996/8.022219
[1] 1.135596
```

– así que algo menos del 13.6%. Sin embargo, esto varía en función de la muestra. Por supuesto, todo el procedimiento puede simularse. Suponemos los mismos valores para las poblaciones respectivas (véase el cuadro 4.4), extraemos dos muestras aleatorias de cada una de estas poblaciones y calculamos la diferencia media y la desviación típica agrupada en cada caso, así como las respectivas Intervalos de condensación del 95%. Todo esto se escribe en una tabla, cuyo principio y final vemos a continuación. cuyo final veremos a continuación:

```
# simulation mean differences between two samples
n1 <- 50
mu1 <- 100
sigma1 <- 10
n2 <- 48
mu2 <- 104.8
sigma2 <- 12.5
trials <- 1000
min.n1 <- 5
min.n2 <- 5
max.n1 <- 100
max.n2 <- 100
random.n <- FALSE
seed <- 9876
set.seed(seed)
# actual simulation
res <- do.call("rbind", lapply(seq_along(1:trials), function(i)
{
# create random sample sizes
if(random.n == TRUE)
{
n1 <- sample(min.n1:max.n1, size=1)
n2 <- sample(min.n2:max.n2, size=1)
}
samp1 <- rnorm(n=n1, mean=mu1, sd=sigma1)
samp2 <- rnorm(n=n2, mean=mu2, sd=sigma2)
xbar1 <- mean(samp1)
sd1 <- sd(samp1)
xbar2 <- mean(samp2)
sd2 <- sd(samp2)
ci.dim.res <- ci.diff.in.means(n1=n1, xbar1=xbar1,
            sd1=sd1, n2=n2, xbar2=xbar2, sd2=sd2,
            equal.var=FALSE)
data.frame(n1, n2, "xbar1"=xbar1, "xbar2"=xbar2, "s1"=sd1, "s2"=sd2,
            ci.dim.res[3,][-6], check.names=FALSE)
}))
rownames(res) <- 1:dim(res)[1]
# results
head(res)
tail(res)
```

```
# show plot and results
res.plot <- plot.CI(res=res, trials=trials)
```

La figura 4.8 muestra el histograma de las diferencias medias simuladas. Las líneas verticales (discontinuas) marcan los límites superior e inferior del intervalo de confianza del 95% en todas las muestras, donde utilizamos la desviación estándar de la muestra bootstrap para calcular los límites superior e inferior. La línea vertical continua marca la diferencia de medias. Como puede observarse, los valores se distribuyen de forma relativamente simétrica en torno a la diferencia media. distribuidos en torno a la diferencia media. Comparemos los resultados de la simulación ...

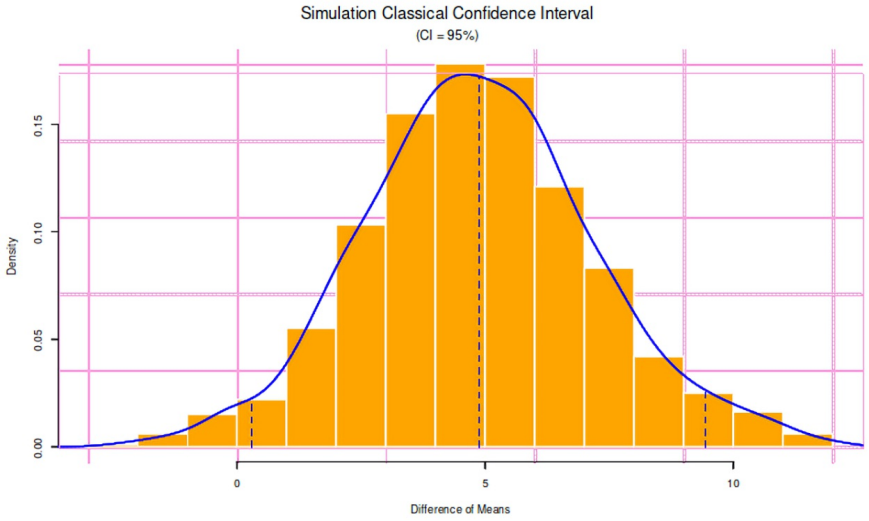


Figura 4.8 *Intervalos de confianza de simulación (diferencias de medias)*

```
> digits <- 3
> cat(paste("\n",round(res.plot$delta.mean,digits),
+           " [",round(res.plot$ci.low,digits),
+           "; ",round(res.plot$ci.up,digits),
+           "] [",round(res.plot$ci.up-res.plot$ci.low,digits),
+           "]\n\n",sep=""))
4.871 [0.302; 9.44] [9.138]
```

... con los valores de una única comparación (casi una muestra de la simulación bootstrap), ...

```
> ci.diff.in.means(n1=n1, xbar1=xbar1, sd1=sd1, n2=n2, xbar2=xbar2,
+ sd2=sd2, equal.var=FALSE)
            N  CI(low) theta CI(up) CI(width) SD   SE
sample(1)   50  97.158 100.0 102.84 5.68      10.0 1.41
sample(2)   48 101.170 104.8 108.43 7.26      12.5 1.80
delta(mean) 98   0.245   4.8   9.35 9.11      NA   2.29
            CI(prob) t    df   var(equal)
sample(1)   0.95     2.01 49.0 NA
sample(2)   0.95     2.01 47.0 NA
delta(mean) 0.95     1.99 89.9 FALSE
```

... los valores coinciden relativamente bien entre sí. Sin embargo, esto no tiene por qué ser así debido a la variabilidad del muestreo. Por lo tanto, recordemos de nuevo -como afirma Neyman (1937)- que la probabilidad de que el valor empírico se encuentre dentro del intervalo de confianza válido para la población es $p = 0$ o $p = 1$. O bien el valor se encuentra dentro del intervalo especificado o bien no – y esto es independiente de que se realicen pruebas o no. Por desgracia, cuando se determina la confianza, nadie sabe exactamente cuál de los $(1 - \alpha)\%$ de los valores se encuentran dentro de la confidencia o no. Sólo cuando se conozcan los valores reales se establezca sin lugar a dudas. Podemos acercarnos relativamente a los

valores reales mediante infinitas repeticiones de muestras o mediante buenas simulaciones, si las condiciones de simulación se corresponden exactamente con las reales. Si las condiciones no se corresponden con las reales, por ejemplo porque no tenemos en cuenta variables de influencia importantes, la simulación no puede ayudarnos, sino que sólo aumenta el error total, porque nos "creemos" la simulación, sobre todo porque corresponde a una gran cantidad de datos.

Como puede verse, con los intervalos de confianza se hace una afirmación sobre la población a partir de repeticiones potencialmente infinitas en condiciones exactamente idénticas y no se hace ninguna afirmación sobre una muestra concreta (aleatoria). Si sólo disponemos de una muestra concreta y se desconocen los valores reales, siempre existe una gran incertidumbre sobre los parámetros de la población, especialmente si no hay replicación. Este razonamiento por sí solo ya debería bastar para una exigencia general de réplicas, lo que no significa que sea fácil. Por desgracia, la práctica demuestra que, obviamente, esto no es suficiente (Kahnemann, 2012; Open Science Collaboration, 2015; Bohannon, 2015). El planteamiento de Neyman-Pearson resulta más claro si tomamos el ejemplo de la gestión de la calidad (variante de caso 1 - piezas de automóvil) (véase el capítulo 4.3.3.4).

> Caso 4.3: Variante de producción de piezas de automóvil
>
> Por ejemplo, queremos tener como máximo un 2% de rechazo. Así que fijamos por adelantado = 2% (*100%* - α*%%*) y ajustamos todos los demás parámetros en consecuencia para un análisis de potencia a priori. Ahora simplemente queremos decidir "sólo" por prueba: Criterio cumplido o no en el nivel. A diferencia de Fisher, a nosotros no nos interesa el valor *p* exacto. Nos gustaría alcanzar una confianza alta, de modo que podamos estar seguros con la probabilidad de *(1-*α*)%* de producir un α*%* máximo de rechazo (= "tasa de falsos descubrimientos") en todas las pruebas, que se clasifica como defectuosa, pero que en realidad no lo es (prueba unilateral debido a consideraciones de contenido). Puesto que nosotros mismos fijamos el nivel, el porcentaje de confianza no es una cantidad empírica, sino que forma parte de las condiciones marco que establecemos antes del examen o de nuestro proceso de producción. El intervalo *(1-*α*)%* corresponde a nuestro intervalo de confianza y, por tanto, a nuestros límites de tolerancia.

A partir de las explicaciones anteriores sobre los intervalos de confianza, queda claro que los niveles α e β no son errores que se producen una vez, sino tasas de error. Se aplica el supuesto de repeticiones infinitas, de modo que por ejecución se puede cometer potencialmente tanto un error α como un error β y producirse en términos porcentuales. En la suma de todas las ejecuciones, los errores singulares se convierten en una tasa de error (Hubbard & Bayarri, 2003; Hubbard, 2004; Hubbard & Lindsay, 2008), que corresponde exactamente a la definición de α y β de Neyman-Pearson. Y ambas pueden visualizarse como distribuciones (ejemplificadas en la Fig. 4.6, p.102, Fig. 4.7, p.103 y Fig. 4.23, p.155, respectivamente).

También debe quedar claro que – véanse las variantes del caso ficticio, cap. 4.3.3.4 – una determinación exacta del nivel α en un contexto de gravedad debe dar lugar inmediatamente a un cálculo de costes completo (véanse cap. 4.3.3.5 o 7.10.2) para no poner en peligro la empresa ni vender productos peligrosos. Desde el punto de vista empresarial, la cuidadosa elección de los parámetros salta inmediatamente a la vista. No nos queda claro si esto también se aplica a los científicos (sociales), ya que aquí la seriedad – es decir, las consecuencias reales de los resultados de la investigación – no se configura de la misma manera que puede hacerse en la economía libre.

> **Tarea 4.4: Dirección de la hipótesis**
>
> Así que tenemos que preguntarnos: ¿Qué es peor? ¿Vender los productos rechazados como si no tuvieran defectos, como en el caso ficticio (nivel β), o identificar los productos sin defectos como defectuosos y no venderlos en absoluto (nivel α)? Invitamos al lector a plantearse varias hipótesis escenarios de su propio campo y aplicar la discusión de las hipótesis a su propio proyecto de investigación. Para ello, cambie el nivel de las tasas de error α y β y comparense entre sí las consecuencias resultantes.

La siguiente simulación R facilita la comprensión de cómo se producen los intervalos de confianza. Creamos k = 100 simulaciones con los parámetros de población μ = 4.5 y σ = 1.7, que son valores elegidos arbitrariamente. Dado que los verdaderos parámetros de población son fijos, no hay error estándar. Esto sólo existe con muestras, ya que allí – al prevalecer condiciones finitas – es igual a CERO:

$$\sigma_{SE} = \frac{\sigma}{\sqrt{N}} \qquad (4.14)$$

Si se aplica un nivel convencional α = 5%, podemos contar cuántos intervalos de confianza cubren la media verdadera conocida y predefinida y cuáles no la cubren. La función `CI.evolve()` de R genera números aleatorios distribuidos normalmente para un número determinado de repeticiones basándose en una entrada, que corresponde a los parámetros de la población y, por tanto, a los valores reales (media, desviación estándar y tamaño de la muestra). Cada repetición simulada corresponde a una réplica en condiciones idénticas. Para cada simulación, se calcula un intervalo de confianza clásico para el valor medio y se almacenan las variables necesarias para ello. Una visualización (véase la Fig. 4.9) de la simulación con `errbar()` del paquete `Hmisc` de R muestra el curso de los intervalos de condensación. La línea horizontal roja marca el valor medio de todas las muestras de la simulación, que se aproxima cada vez más al valor poblacional. Esto se marca con la línea horizontal naranja. Los puntos rojos indican el valor medio respectivo por simulación. Las líneas verticales corresponden a un intervalo de confianza. Las pequeñas líneas horizontales indican el principio y el final del intervalo de confianza. Si el intervalo de confianza no incluye la media global (es decir, el parámetro poblacional), entra en el intervalo de rechazo denotado por el valor α. El valor del intervalo de confianza corresponde a *1* - α. Por término medio, cabe esperar tantos intervalos de confianza en la área α como el número de réplicas multiplicado por el nivel α . Cuanto mayor sea la muestra total más precisas serán las cifras y más se acercarán a los valores de la población. El siguiente código R implementa esto (`ptII_quan_classicstats_N-P_confint-errorbars.r`) y muestra el resultado en la Figura 4.9.

```
trials <- 100
pop.mean <- 4.5
pop.sd <- 1.7
sim.res <- CI.evolve(N=30, trials=trials, pop.mean=pop.mean,
pop.sd=pop.sd, prob=(prob <- 0.95))
```

Por lo tanto, existen desviaciones cuando el intervalo completo no encierra la media, que en la suma de todos estos valores corresponde aproximadamente al nivel (100-95)% = 5% de la población para el que el valor verdadero se encuentra fuera del intervalo de confianza, aunque represente una muestra *legítima* y, por lo tanto, auténtica de la población seleccionada. La salida R muestra el estadísticas descriptivas.

```
# inspect data
str(sim.res)
head(sim.res$simulation)
tail(sim.res$simulation)
```

```
sim.res[-1]
# descriptive statistics
psych:::describe(sim.res$simulation)
```

Si comprobamos la precisión con CI.cover.mean() ya con el nivel de confianza elegido del 95%, veremos que es más o menos correcta. Con 100 simulaciones esperamos (100-95) * 100% = 5% de intervalos de confianza que no cubren la media global (población).

```
# repeat for 100 trials R-Code
trials <- 100
sim.res1 <- CI.evolve(N=30, trials=trials, pop.mean=pop.mean,
           pop.sd=pop.sd, seed=9876)
CI.cover.mean(sim.res=sim.res1)
```

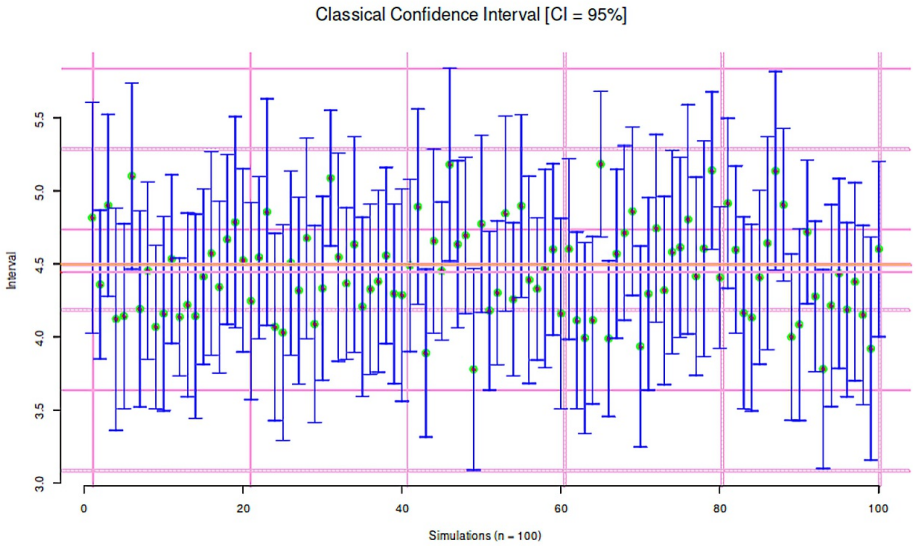


Figura 4.9. *Simulación de intervalos de confianza*

Aquí hay 7 intervalos de confianza. Si repetimos todo para 1000 simulaciones cambiando el valor inicial de los ensayos con los mismos parámetros de población y la misma salida de números aleatorios utilizando `seed()`, ahora obtenemos 46 valores, es decir, el 4.6%.

Así es exactamente como debería desarrollarse. Esto se ajusta bastante bien a las expectativas y debería aumentar nuestra confianza a la hora de traducir supuestos teóricos en código R práctico con números. Con 10 000 repeticiones, el resultado es de 4.87%. La media empírica de las simulaciones es ahora idéntica con tres decimales al valor poblacional $\mu = 4.5$. La desviación estándar *s* del ejemplo R corresponde al 99.24% del valor poblacional $\sigma = 1.7$. Sin embargo, si realizamos el mismo ejercicio una sola vez – como suele ocurrir en las ciencias sociales (véase la Fig. 4.10) – existe una clara diferencia entre la media empírica y la media poblacional. Recordemos que aquí, en la simulación, conocemos los verdaderos parámetros de la población. En la práctica, sin embargo, se desconocen y pueden desviarse hacia arriba o hacia abajo en mayor o menor medida.

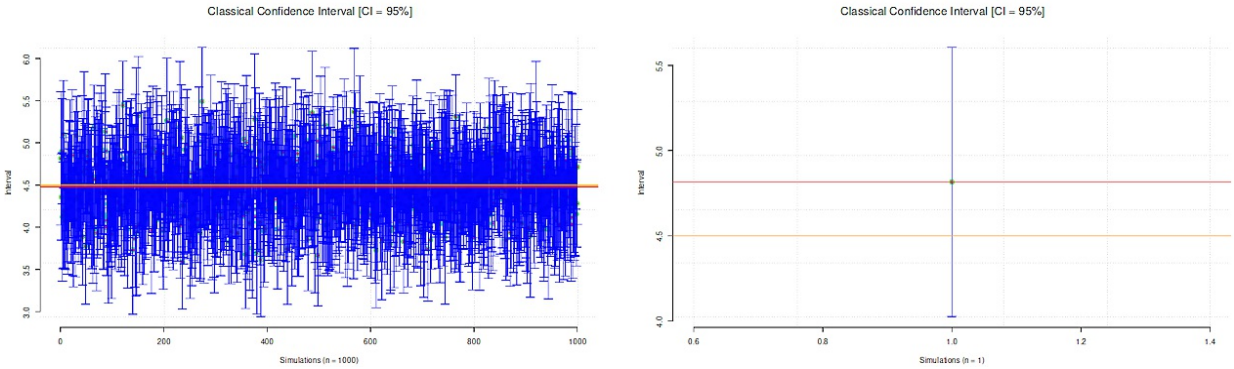


Figura 4.10. *Intervalos de confianza de la simulación (pruebas únicas frente a repetidas)*

```
# repeat for 1 trials
trials <- 1
sim.res1 <- CI.evolve(N=30, trials=trials, pop.mean=pop.mean,
            pop.sd=pop.sd, seed=9876)
CI.cover.mean(sim.res=sim.res1)
```

Puesto que aquí conocemos la verdadera media de la población, la figura 4.9 muestra claramente que podemos examinar cada una de las muestras al nivel α para ver si el intervalo empírico incluye o no el parámetro poblacional y, por tanto, si debe rechazarse o no H_0. El cambio de pruebas de significación a intervalos de confianza no utiliza realmente ninguna información nueva de la muestra actual que no se haya utilizado ya (véase el capítulo 4.3.3.6), a saber, x_1, s_1,n_1 así como x_2, s_2,n_2 y el nivel α. La misma información entra en la prueba *t* de dos muestras (véanse las fórmulas anteriores), para seguir con el ejemplo anterior de la diferencia de medias entre dos muestras independientes con varianzas desiguales.

Aquí también podemos calcular las confidencias, de nuevo empíricamente mediante simulación. Para ello, utilizamos la función de R `sim.ttest()` con los parámetros para las dos muestras ($n_1 = n_2 = 30$, $\mu_1 = 4.7$, $\sigma_1 = 1.8$, $\mu_2 = 4.4$, $\sigma_2 = 1.7$). Generamos dos distribuciones aleatorias basadas en los parámetros poblacionales especificados. Se comprueban entre sí con una prueba t de Weelch de dos muestras utilizando `t.test()` (muestras independientes, varianzas desiguales, prueba de dos caras). Las estadísticas calculadas en el proceso (valor t, valor p al nivel habitual, diferencia de medias, etc.) se guardan, así como la potencia del efecto d de Cohen (Cohen, 1992).

```
# initial values
alpha <- 0.05
seed <- 9876
n1 <- 30
mu1 <- 4.7
sigma1 <- 1.8
n2 <- 30
mu2 <- 4.67#4.4
sigma2 <- 1.7
trials <- 1000
ttest.res <- sim.ttest(n1=n1, mu1=mu1, sigma1=sigma1, n2=n2, mu2=mu2,
             sigma2=sigma2, trials=trials,
             seed=runif(1)*1000)
str(ttest.res)
head(ttest.res)
# plot p- and t-values (histogram and log scaled)
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l", mfrow=c(2,2))
hist(ttest.res[,"t"], pre.plot=grid(), col="darkred", border="white",
             xlab="t-values", prob=TRUE, main=NA)
lines(density(ttest.res[,"t"]), col="steelblue",lwd=2, lty=1)
hist(ttest.res[,"p"], pre.plot=grid(), col="orange", border="skyblue",
```

```
            xlab="p-values", prob=TRUE, main=NA)
lines(density(ttest.res[,"p"]), col="steelblue",lwd=2, lty=1)
plot(log(sort(ttest.res[,"p"])), col="darkred", type="l", bty="n",
            pre.plot=grid(), ylab="log(p-value)")
plot(as.brob(sort(ttest.res[,"t"])), col="darkred", type="l", bty="n",
            pre.plot=grid(), ylab="log(t-value)")
mtext(text="Simulation t-test (t- and p-values)", outer=TRUE,
            line=-2, cex=1.5, side=3)
```

A continuación examinamos los estadísticos descriptivos y los histogramas (valores t, valores p, diferencias de medias, d de Cohen). Como puede verse en las figuras los valores t y las diferencias medias se distribuyen simétricamente (véanse las figuras 4.11 y 4.13). En cambio, los valores p no lo son, lo que tiene consecuencias a la hora de calcular las confidencias, desviaciones estándares etc., es decir, todo lo que requiere simetría. Lo mismo ocurre con las correlaciones, que también tienen una distribución sesgada (palabra clave: Transformación z de Fisher según Fisher, 1915, 1921). El siguiente código R es útil (`ptII_quan_classicstats_N-P_confint_p-t-value.r`).

```
# p-values lower than alpha R-Code
sum(ttest.res[,"p"] < alpha)/trials
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l", mfrow=c(1,2))

# plot first few p-values
plot(ttest.res[1:100,"p"], pre.plot=grid(), type="b", col="red",
lty=1, bty="l", xlab="trial", ylab="p-value", main="")
abline(h=c(0.01,0.05), col="blue", lty=2, lwd=1)

# plot d
plot(ttest.res[1:100,"d"], pre.plot=grid(), type="b", col="dark green",
lty=1, bty="l", xlab="trial", ylab="Cohen's d", main="")
abline(h=c(0.4,-0.4), col="red", lty=2, lwd=1)
mtext("Simulations (process)", outer=TRUE, line=-2, cex=1.5, side=3)
```

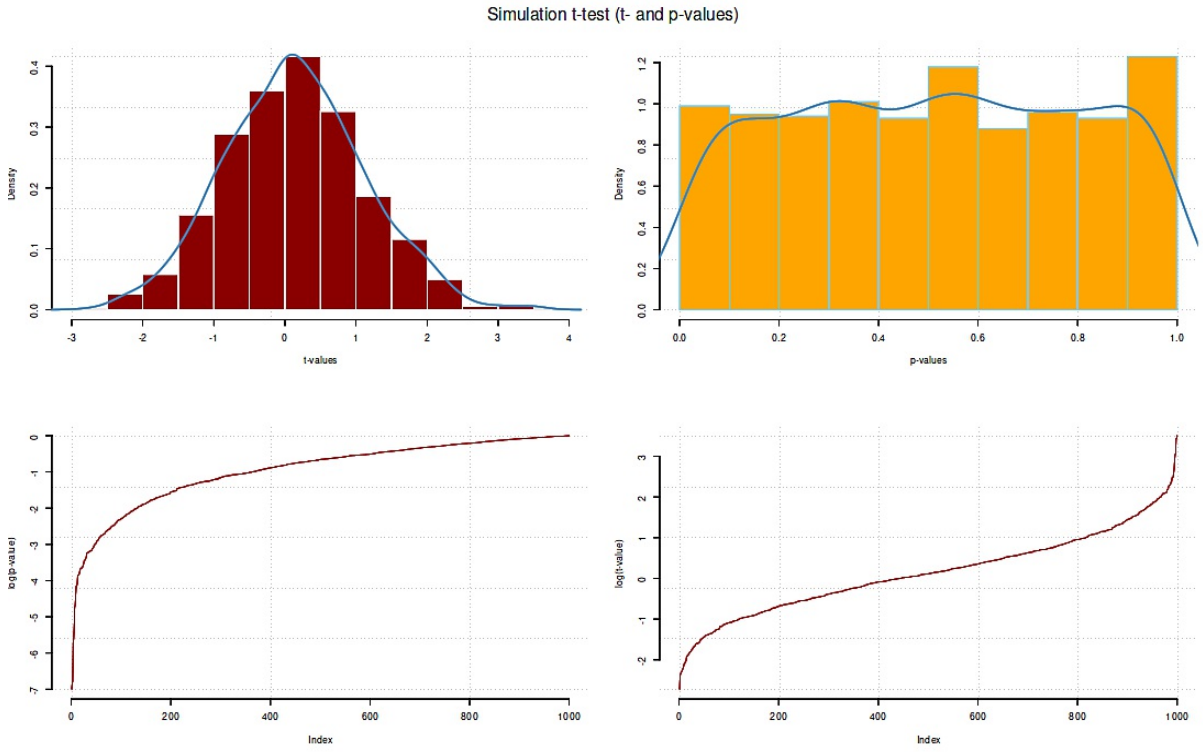


Figura 4.11. *Simulación de prueba t (valores t y p)*

Si nos interesan las condensaciones de las diferencias de medias, podemos calcularlas a partir de la media y el error estándar o a partir de los cuantiles utilizando `quantile()`. Con muestras de pequeño tamaño, estos valores son propensos a errores, es decir, inexactos, pero tienen una base empírica. Un pequeño ejemplo en R lo demuestra. Generamos 1001 valores a partir de una distribución t con infinitos grados de libertad y posteriormente calculamos los cuantiles a partir de probabilidades, es decir, porcentajes de área bajo la curva. Comparamos el resultado con la función exacta `qt()`. El resultado nos da la sobreestimación o subestimación en comparación con los valores exactos de la distribución. Esto nos permite formarnos una idea de las diferencias entre el empirismo (sujeto a error) y la expectativa exacta de la teoría. Si volviéramos a aumentar el tamaño de la muestra, la relación entre los valores empíricos y los teóricos mejoraría sucesivamente. Esta sería una tarea R para los lectores. A continuación, también podrían trazarse las curvas para diferentes tamaños de muestra.

```
> # show how quantile() works
> N <- 1001
> probs <- c(0, 0.025, 0.05, 0.25, 0.5, 0.75, 0.95, 0.975, 1)
> x <- rt(N, df=Inf)
> q1 <- quantile(x, probs=probs)
> q2 <- qt(probs, df=Inf, lower.tail=TRUE)
> q1
 0%      2.5%    5%      25%     50%    75%    95%    97.5%  100%
-2.9550 -1.8533 -1.6260 -0.6664 -0.0497 0.7147 1.6678 1.8403 3.4669
> q2
[1] -Inf -1.960 -1.645 -0.674 0.000 0.674 1.645 1.960 Inf
> q1/q2*100
0%  2.5% 5%   25%   50% 75%   95%   97.5% 100%
0.0 94.6 98.9 98.8 -Inf 106.0 101.4 93.9  0.0
```

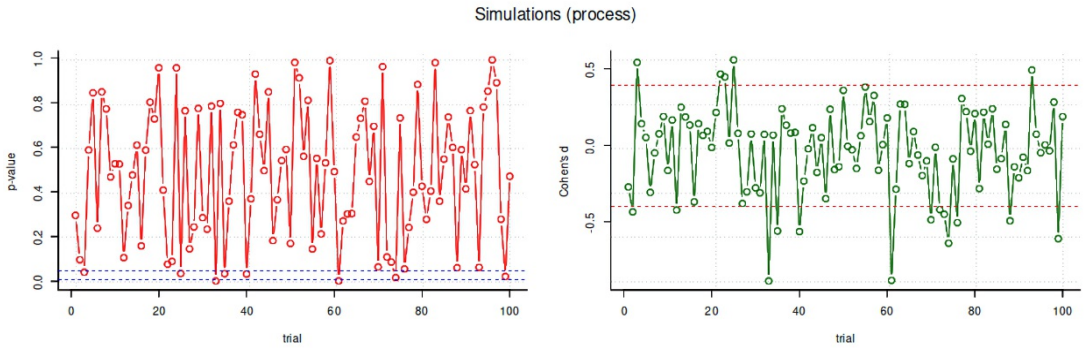


Figura 4.12. *Prueba t de simulación (Sección del curso de la valores p y d de Cohen)*

Si aplicamos nuestro conocimiento de la imprecisión empírica a las diferencias medias generadas anteriormente, obtenemos una distribución de diferencias medias. Podemos visualizar el resultado con la Figura 4.13. Las líneas verticales discontinuas en los extremos de la distribución indican la media y los límites de confianza del 5% y el 95%.

```
> # confint
> prob <- 0.95
> alpha <- 1-prob
> cohensd <- ttest.res[,"d"]
> probs <- c(0, 0.025, 0.05, 0.25, 0.5, 0.75, 0.95, 0.975, 1)
> quantile(cohensd, probs=probs)
 0%      2.5%    5%      25%     50%    75%     95%    97.5%  100%
-0.9049 -0.5405 -0.4741 -0.1945 -0.0293 0.1406  0.3704 0.4507 0.7011
> summary(cohensd)
 Min.   1st Qu. Median Mean  3rd Qu. Max.
-0.905 -0.194  -0.029 -0.036 0.141   0.701
>
> plot.d.sim(cohensd=cohensd)
```

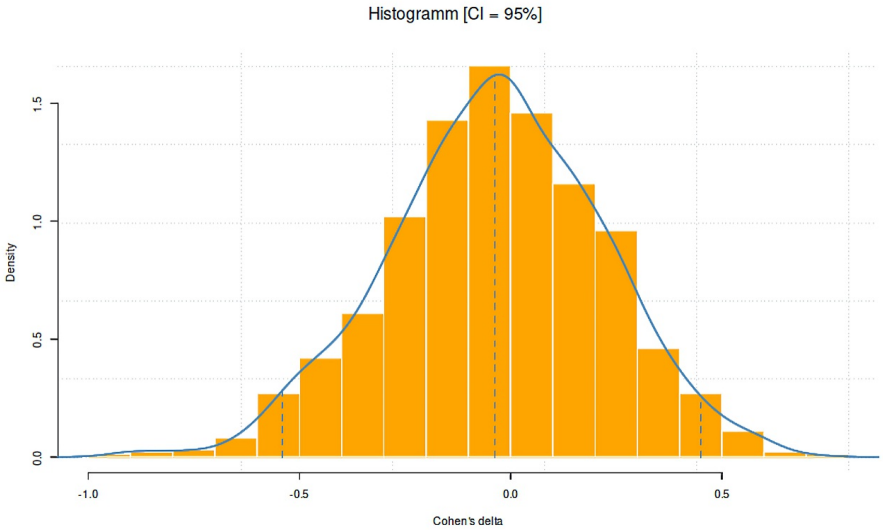


Figura 4.13. *Simulación de diferencias de medias*

El siguiente código R muestra los resultados bastante comparables de simulaciones manuales con código propio, con la ayuda de `boot()` del paquete R `boot` o de modo bayesiano con `bayesboot()` del paquete R `bayesboot`. Simula las diferencias de medias de dos muestras distribuidas normalmente, tanto en forma de valores *t* como de *d* de Cohen (desviaciones estándares de muestras agrupadas). Primero el caso manual `(ptII_quan_classicstats_N-P_confint_bayesboot.r);` imprimimos el código R sólo para el cálculo de la d de Cohen:

```
# number of replications / bootstrap samples
repli <- 1e+3
#seed
seed <- 56569
set.seed(seed)
# boot t-value and Cohens'd between two samples
N1 <- 43
N2 <- 45
mu1 <- 1.89
mu2 <- 1.5
s1 <- 0.5
s2 <- 0.4
#create sample
samp1 <- rnorm(N1, mean=mu1, sd=s1)
samp2 <- rnorm(N2, mean=mu2, sd=s2)
samplevalues <- c(samp1, samp2)
group <- c(rep(1,N1),rep(2,N2))
```

Y ahora con `boot()`:

```
# boot cohens's d 's pooled' version
boot.d <- function(dats, ids)
{
# calculate delta based on sample2 minus sample1
return( cohensd( dats[ids[group==2]], dats[ids[group==1]] )[2] )
}
set.seed(seed)
b.d.res <- boot(samplevalues, boot.d, strata=group, R=repli)
b.d.res
boot.ci(b.d.res)
plot(b.d.res, jack=TRUE)
boot.ci(b.d.res, conf=c(0.5,0.89,0.95),
type=c("norm", "basic", "perc", "bca"))
```

Y al final con `bayesboot()` (véase figura 4.14):

```
# bayesboot cohen's d 's pooled' version
set.seed(seed)
bb.mean1 <- bayesboot(samp1, mean, R=repli)
bb.mean2 <- bayesboot(samp2, mean, R=repli)
bb.sd1 <- bayesboot(samp1, sd, R=repli)
bb.sd2 <- bayesboot(samp2, sd, R=repli)
bb.s.pooled <- as.bayesboot(sqrt(((N1-1)*bb.sd1^2 +
(N2-1)*bb.sd2^2) / (N1+N2-2)))
# bayesboot cohen's d 's pooled' version
bboot.meandiff <- as.bayesboot(bb.mean1-bb.mean2)
bboot.d.res <- as.bayesboot( bboot.meandiff/bb.s.pooled )
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l")
plot(bboot.d.res, showCurve=TRUE, showMode=TRUE, pre.plot=grid())
mtext("Bayes Bootstrap", outer=TRUE, line=-1, cex=1.5, side=3)
```

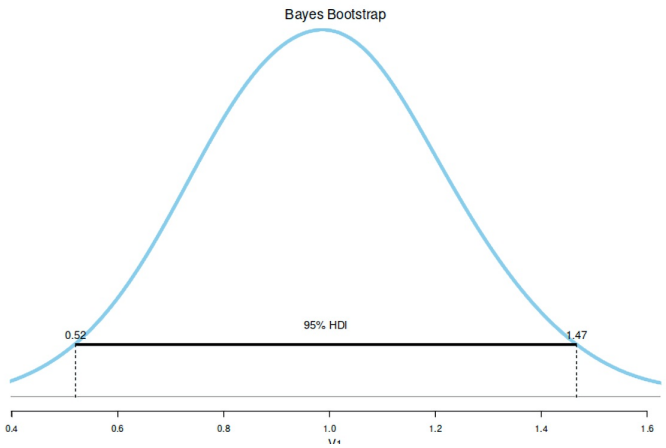


Figura 4.14. *Simulación de confidencias con* `bayesboot()`

¿Qué muestran los resultados en términos de confianza? La interpretación se deja a los lectores como tarea. Deben consultarse las páginas del manual de las respectivas funciones R. Puede comparar los resultados (el mensaje de advertencia emitido no es relevante aquí) con

```
> boot.ci(b.d.res)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates
CALL :
boot.ci(boot.out = b.d.res)
Intervals :
Level  Normal           Basic
95%    ( 0.545, 1.357 ) ( 0.552, 1.360 )
Level  Percentile       BCa
95%    ( 0.552, 1.360 ) ( 0.544, 1.353 )
Calculations and Intervals on Original Scale
Mensaje de advertencia:
In boot.ci(b.d.res) :
Bootstrap-Varianzen für studentisierte Intervalle benötigt
```

O también

```
> summary(bboot.d.res)
Bayesian bootstrap
```

```
Number of posterior draws: 1000
Summary of the posterior (with 95% Highest Density Intervals):
statistic mean  sd   hdi.low hdi.high
V1        0.991 0.22 0.558   1.41
Quantiles:
statistic q2.5% q25%  median q75% q97.5%
V1        0.575 0.836 0.985  1.13 1.43
Call:
```

En el script de R existen otras posibilidades de comparación tabular. Con la variante bayesiana se pueden examinar directamente las probabilidades relativas a los valores interesantes. Por ejemplo, podemos preguntarnos cuántos puntos porcentuales de la distribución posterior se sitúan por encima de un valor *t* superior a 2.

```
> # direct comparisons Bayes posterior distribution
> mean( bb.tv.res > 7)
[1] 0.015
> mean( bb.tv.res > 7 & bb.tv.res < 9.5)
[1] 0.015
> mean( bb.d.res > 0.8)
[1] 0.81
> mean( bb.d.res > 0.5)
[1] 0.988
```

O podríamos preguntar cuántos porcentajes de la distribución posterior están por encima de una *d* de Cohen de 0.2 y por debajo de 0.8 o superior a 0.8 e inferior a 1.26:

```
mean( bb.d.res > 0.2 & bb.d.res < 0.8)
[1] 0.19
> mean( bboot.d.res > 0.8 & bboot.d.res < 1.26)
[1] 0.7
```

4.3.4 Pruebas y estimaciones: La relación entre confianza y valor p

La conexión entre las confidencias y una prueba estadística – como la validación estadística de β-pesos frente a CERO en el curso de un análisis de regresión – no es tan sencilla como parece a primera vista (Thulin, 2014; Thulin & Zwanzig, 2017). Recapitulemos primero la aparición respectiva sobre la base de las explicaciones anteriores -- las confidencias resultan de la adición/subtracción del parámetro de interés y su error estándar multiplicado por el intervalo de confianza seleccionado. Un valor *t* para una prueba de significación (frente a cero) resulta del cociente del parámetro estimado y su error estándar. Así, comúnmente en los modelos lineales, los coeficientes individuales están asegurados contra CERO. Esto corresponde a la fórmula general de la prueba *t*, cuya variable de prueba distribuida *t* (un cociente) representa un valor normalizado a desviaciones estándar. En un sentido más amplio, la confianza consiste en la *suma (diferencia)* del parámetro y el producto del error estándar por el intervalo de confianza H_0

```
2*qnorm(1-alpha/2, lower=TRUE)
```

mientras que el valor *t* es el producto (cociente) del parámetro y el error estándar, que a su vez se localiza en la distribución H_0. Un ejemplo de la página de ayuda (en R) de `lm()` muestra esto con un conjunto de datos de Dobson (1990, p.9) sobre el peso de las plantas (grupo de control, grupo de tratamiento):

```
ctl <- c(4.17,5.58,5.18,6.11,4.50,4.61,5.17,4.53,5.33,5.14)
trt <- c(4.81,4.17,4.41,3.59,5.87,3.83,6.03,4.89,4.32,4.69)
group <- gl(2, 10, 20, labels = c("control","treatment"))
weight <- c(ctl, trt)
lm.fit <- lm(weight ~ group)
```

En primer lugar, la salida de la estimación del modelo lineal, que incluye los valores de los parámetros y sus sus errores estándar de estimación y los valores *t* de la regresión:

```
> summary(lm.fit)
Call:
lm(formula = weight ~ group)
Residuals:
 Min     1Q     Median 3Q     Max
-1.0710 -0.4937 0.0685 0.2462 1.3690
Coefficients:
                Estimate Std.   Error t value Pr(>|t|)
(Intercept)     5.032    0.220  22.85 9.5e-15 ***
grouptreatment -0.371    0.311  -1.19 0.25
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.696 on 18 degrees of freedom
Multiple R-squared: 0.0731, Adjusted R-squared: 0.0216
F-statistic: 1.42 on 1 and 18 DF, p-value: 0.249
```

Ahora calculamos a mano el valor *t* para la comparación del grupo de control frente al de tratamiento. Lo comparamos con el de la salida R. Para evitar errores de redondeo, comprobamos la diferencia con `all.equal()`:

```
> # t-value 'grouptreatment'
> all.equal(-0.3710/0.3114, summary(lm.fit)$coef[2,3])
[1] "Mean relative difference: 0.000112"
```

Como era de esperar, la diferencia está en el rango del error de redondeo. Ahora sigue la parte complementaria, en la que se calcula un intervalo de confianza clásico del 95% para el parámetro estimado.

```
> # CI
> alpha <- 0.05
> fak <- qnorm(1-alpha/2, lower=TRUE)
> # CI.low
> summary(lm.fit)$coef[2,1] - fak * summary(lm.fit)$coef[2,2]
[1] -0.981
> # CI.up
> summary(lm.fit)$coef[2,1] + fak * summary(lm.fit)$coef[2,2]
[1] 0.239
> # true mean difference
> mu1 - mu2
[1] 0.39
```

Como puede verse, el valor *t* permite hacer afirmaciones sobre la localización del parámetro estimado en relación con una distribución, mientras que el intervalo de confianza clásico hace afirmaciones sobre el intervalo en el que, según la teoría de Neyman-Pearson, son de esperar valores para la comparación de grupos estimada. Se trata de dos tipos diferentes de declaraciones que se basan en las mismas estadísticas y, por tanto, son equivalentes.

Sin embargo, esto no tiene por qué ser siempre así. Como señala Thulin (2014), el valor *p* y el intervalo de confianza clásico son equivalentes si se basan *en el mismo estadístico*: „We can interpret the p-value as the smallest value of for which the null value of the parameter would be included in the 1−confidence interval“ (Thulin, 2015-05-28). El método de construcción de intervalos aquí descrito se basa en el uso de un estadístico relacionado con el parámetro θ desconocido en que nos interesamos. Además, existen otros intervalos basados en algoritmos de minimización. Aquí, la longitud del intervalo se minimiza con respecto a la variable aleatoria investigada. Estos intervalos ya no son equivalentes al valor *p* o a la prueba estadística. En principio, la equivalencia se pierde si los estadísticos para calcular el intervalo de confianza y los de la prueba estadística (valor p) son diferentes (Fay, 2010). Pero incluso si la equivalencia existe, no es necesariamente significativa y deseable, como señala (Thulin, 2015-05-28), ya que „sometimes intervals and tests have somewhat conflicting goals. We want short intervals and tests with high power, but the shortest interval does not always correspond to the test with the highest power.“

Por lo tanto, es necesario examinar exactamente qué información se utiliza en los cálculos. Esto es así independientemente de que las interpretaciones del contenido difieran en función de la combinación de la información disponible. Desde el punto de vista de la teoría de la información, no hay ganancia de conocimiento si la misma información se utiliza de forma equivalente. De acuerdo con la distinción anterior, se puede generar una prueba de hipótesis a partir del intervalo de confianza mediante reformulación, lo que en última instancia apunta a la identidad mayoritariamente verdadera del intervalo de confianza y el texto de significación clásico. Según Tschirk (2014, p.82), la dualidad de estimación y prueba se aplica "cuando la desviación estándar de la estimación puntual es la misma que la de la variable de prueba en caso de una hipótesis nula correcta."

Por supuesto, existen argumentos en contra de este punto de vista, como el hecho de que los intervalos de confianza "no requieren hipótesis a priori y, por tanto, [evitan] la comprobación de hipótesis nulas triviales" (Brandstätter, 1999, p.2). Los intervalos de confianza muestran la precisión de las estimaciones a través de su amplitud o estrechez. Hay que tener en cuenta "como de costumbre" la arbitrariedad común qua convención en la elección del *p%* de confianza, que es idéntico al (1-α)%.

En resumen, aunque a menudo se basan en la misma información que las pruebas de significación, las confidencias pueden ofrecer una visión más amplia de los datos, como las réplicas, las visualizaciones o los metaanálisis (Brandstätter, 1999, p.14). Como siempre la elección del nivel (1-α)% debe ser cuidadosa, ya que viene determinado por la fórmula común

$$CI = \bar{x} \pm z \cdot \frac{\sigma}{\sqrt{n}} \qquad (4.15)$$

que configura directamente el aspecto del intervalo de confianza y rara vez se pone en duda. El valor medio *x*, el valor *z* del percentil en cuestión (1 - α), la desviación estándar σ y el tamaño de la muestra *n* entran en la fórmula anterior. Esta circunstancia de no cuestionamiento tiene en común el clásico intervalo de confianza con las barreras de significancia. El intervalo de confianza pertenece al campo de la estimación y no de la prueba, porque no prueba hipótesis sino que demuestra el comportamiento de los parámetros en determinadas condiciones. Como señalan muchos autores, las confidencias son más fáciles de entender, sobre todo para los principiantes, que las pruebas de significación y las hipótesis asociadas. Sin embargo, en caso de que las pruebas de significación y los intervalos de confianza se basen en estadísticas diferentes y por eso pautas diferentes de procesamiento de la información o persigan objetivos diferentes, la equivalencia descrita ya no existe o contiene objetivos contradictorios. En la práctica, surge el problema de que la desviación estándar no puede calcularse sin más para cada parámetro. Por tanto, las confidencias pueden estimarse mediante simulaciones.

4.3.5 Excursus – Simulaciones

En principio, las simulaciones pueden realizarse elegantemente con las funciones R del paquete `boot()`. La función principal `boot()` permite generar estadísticas de interés utilizando la simulación `bootstrap` (Efron, 1979; Efron & Tibshirani, 1993) mediante funciones especialmente definidas. El *bootstrap* pertenece al grupo de técnicas de remuestreo, es decir, de repetición de muestras (Yu, 2003) e incluye muchas variantes diferentes. La *prueba de aleatorización* (prueba de permutación) se remonta a Fisher (1935/1973) y a su experimento del té (véase el Cap. 4.3.2.1 o 6.13.4), que corresponde a *sacar sin volver a poner en su sitio* (bolas de la urna). A diferencia del bootstrap, al *sacar con volver a poner en su sitio* el número de posibilidades y, por tanto, las respectivas probabilidades relativas disminuyen con cada nueva simulación. Se pueden excluir varios sorteos. Esto da lugar a diferentes formas de distribución. Otros enfoques de simulación incluyen el *jack-knife*, en el que ciertos puntos de datos se omiten alternativamente para el cálculo de las estadísticas de interés. Esto se remonta a Quenouille (1949) y fue desarrollado por Tukey (1958). La *validación cruzada* (cross validation) divide aleatoriamente una muestra en subgrupos. Los resultados de los exámenes se comparan entre sí. Esto es similar a la fiabilidad por mitades (split-half) de la teoría clásica de

los tests. La validación cruzada también se ocupa de la cuestión de la coherencia y la fiabilidad de los resultados. Este enfoque puede complicarse de cualquier manera (Yu, 2003).

El bootstrap se utiliza cuando se desconoce la distribución teórica de los estadísticos de interés (por ejemplo, coeficientes de regresión, diferencias de medias, correlaciones, ...), pero se considera relevante. Lo que no existe se puede simular, al menos hasta cierto punto, basándose en el *teorema del límite central* (Pólya, 1920, véase el capítulo 4.3.9.3). Se distingue entre *bootstrap paramétrico y no paramétrico*. En el caso del bootstrap paramétrico, las simulaciones se realizan a partir de un modelo paramétrico cuyo tipo de distribución se conoce – salvo los parámetros de distribución, mientras que en el bootstrap no paramétrico los datos empíricos constituyen la base de todos los cálculos posteriores. En ambos casos, la incertidumbre radica en que la verdadera distribución de los estadísticos (parámetros) de interés sigue siendo desconocida a pesar de la simulación y por eso la transferibilidad a la realidad no está garantizada. No es casualidad que el término bootstrap, en referencia a la historia del barón Münchhausen, signifique "salir del pantano por sus propios pelos" o "por los cordones de sus propias botas" (del inglés "boot" = bota y "straps" = cordones). Desde el punto de vista analítico, esto demuestra que el bootstrap carece de anclaje en el mundo real y toma lo que está disponible, a menudo sólo unos pocos datos empíricos de una pequeña muestra. La simulación, sin embargo, es una aproximación plausible que puede estar bien justificada en muchas circunstancias. Las subvariantes comunes del bootstrap no paramétrico son el *remuestreo de residuos*, el *bootstrap bayesiano* (Rubin, 1981) o el *remuestreo de casos (completos)*:

- El *remuestreo de residuos* consiste en añadir a las variables dependientes residuos extraídos aleatoriamente de un modelo estimado, creando así nuevas variables dependientes artificiales (datos). Este nuevo conjunto de datos se estima de nuevo con el mismo modelo estadístico y los estadísticos relevantes se almacenan por ejecución para generar las respectivas distribuciones de interés y derivar de ellas, por ejemplo, confianzas para obtener valores plausibles para estimar valores de población.
- El *bootstrap bayesiano* interpreta y genera los conjuntos de datos simulados según la estadística bayesiana reponderando los conjuntos de datos precedentes originales.
- Asimismo, se pueden *volver a muestrear casos* (enteros) o casos subdivididos según criterios. Para más información, consulte el literatura sobre *simulaciones Monte Carlo* (Simon, 1997).

Los problemas con el bootstrap no paramétrico siempre surgen cuando el tamaño de la muestra es pequeño o – más en general – cuando el modo en que se realiza la simulación no representa correctamente la distribución de los parámetros críticos. Por otra parte los problemas con el bootstrap paramétrico surgen cuando el modelo paramétrico elegido no representa correctamente el modelo estadístico en discusión. En ambos casos, el problema radica en un modelo defectuoso. Si, por el contrario, el modelo paramétrico elegido es correcto, basta con tamaños de muestra más pequeños que con el bootstrap no paramétrico, ya que la alineación basal puede ser más estricta. Sin embargo, dada la potencia informática actual, el tamaño de la muestra parece ser el menor de los problemas. Recuerde, no obstante que los modelos siempre son erróneos en términos absolutos, pero pueden ser útiles para trazar mapas y aproximarse a la realidad.

Las aplicaciones del bootstrap son muy diversas. Por ejemplo, se pueden generar intervalos de confianza sin un conocimiento exacto de una distribución, o se puede tomar un modelo lineal empírico estimado como base de una simulación para evaluar el poder predictivo y las limitaciones de un modelo investigado.

Los modelos también pueden probarse entre sí. Ejemplos en R los ofrecen Fox y Weisberg (2018) para modelos lineales y Sánchez-Espigares y Ocaña (2009- 07) para modelos jerárquico-lineales. Una alternativa a la función `boot()` es denotar una función separada que calcule la estadística de interés y escriba un algoritmo que ejecuta el cálculo y almacenamiento repetidos de la variables de interés. La función `sample()` permite sacar con o sin retroceso. Con `replicate(bootrep, FUN)` también se puede simular una distribución bootstrap. Aquí la variable `bootrep` representa el número de repeticiones de bootstrap y `FUN` el número de repeticiones de bootstrap para calcular la estadística crítica.

Para comparar, se simula la diferencia de medias (véase la Fig. 4.15) del ejemplo anterior (`ptII_quan_classicstats_N-P_simulation.r`):

```
seed <- 9876
set.seed(seed)
sim <- function(N1, N2, mu1, mu2, sd1, sd2)
{
x1 <- rnorm(N1, mean=mu1, sd=sd1)
x2 <- rnorm(N2, mean=mu2, sd=sd2)
mean(x1) - mean(x2)
}
boot.means <- replicate(trials, sim(N1=N1, N2=N2, mu1=mu1, mu2=mu2,
+          sd1=sd1, sd2=sd2))
summary(boot.means)
# plot
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l")
hist(boot.means, prob=TRUE, pre.plot=grid(), col="steelblue",
+          border="white", main="", xlab="Mean Differences", ylab="Density")
lines(density(boot.means), type="l", col="red", lwd=2)
```

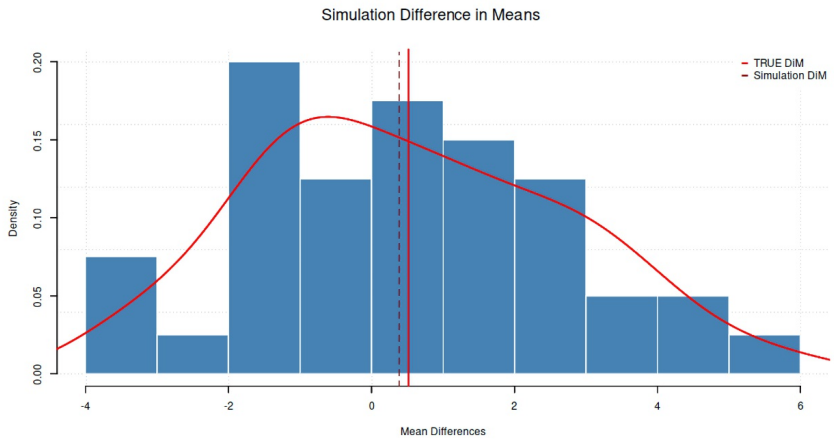


Figura 4.15. *Simulación – Diferencia de medias*

A continuación se compara la diferencia de medias de la población real con la de la simulación:

```
abline(v=mu1-mu2, col="darkred", lty=2, lwd=2)
abline(v=mean(boot.means), col="red", lty=1, lwd=2)
legend("topright", legend=c("TRUE DiM", "Simulation DiM"),
+ pch="---", pt.cex=2.5,col=c("red","darkred"),
+ bty="o", bg="white", box.lty=1, box.col="white")
mtext("Simulation Difference in Means", outer=TRUE,
+ line=-2, cex=1.5, side=3)
```

Y luego todo en cifras:

```
> # true mean difference
> mu1 - mu2
[1] 0.39
> # mean simulation
> mean(boot.means)
[1] 0.518
> # ratio
> abs(1-(mu1-mu2)/mean(boot.means))
[1] 0.248
```

Como puede verse, la simulación funciona. Con 100 simulaciones realizadas, la desviación del valor real es de aproximadamente un 5%. Con 1 000 000 de simulaciones, la desviación es sólo del 0,23% y tiende cada vez más a cero a medida que crece la muestra:

```
> set.seed(seed)
> trials <- 1e+6
> boot.means1 <- replicate(trials, sim(N1=N1, N2=N2, mu1=mu1, mu2=mu2,
+ sd1=sd1, sd2=sd2))
> abs(1-(mu1-mu2)/mean(boot.means1))
[1] 0.00266
```

Del mismo modo, los cálculos para la determinación de las condensaciones pueden realizarse mediante simulación para cualquier otra estadística de interés. otras estadísticas de interés.

4.3.5.1 *Caso práctico: Simulación de imágenes de fútbol*

Probamos las posibilidades de R para la simulación con un ejemplo práctico cotidiano interesante para padres con hijos. ¿Cuánto tenemos que invertir para que nuestro hijo tenga su álbum de fútbol lleno de cromos, sólo porque en la tienda de descuento "de la esquina" los tienen?

La situación inicial es que la tienda de descuento vende cromos de fútbol con fotos de coleccionista de los jugadores de la primera liga nacional de fútbol. Además, se puede comprar un cuadernillo para pegarlas. Llenar el álbum es importante para los niños, por supuesto. Por eso examinamos la opinión de los padres que van de compras con sus hijos. Es cierto en la tienda de descuento que por cada compra de 10 Euros se obtiene un sobre con $n = 5$ cartas, que son siempre cartas diferentes dentro de un mismo sobre. Los sobres también pueden comprarse directamente a 0,70 Euros cada uno, aunque no hay garantía de que no se acumulen imágenes duplicadas. Algunos dependientes de las cajas son muy amables y a veces regalan más paquetes de los que indica la publicidad oficial de las tiendas de descuento. A veces no te las dan e incluso tienes que pedirlas explícitamente. Y, por supuesto, es útil ir de compras con un niño. Ignoramos estas sutilezas por ahora, ya que serían variables adicionales para un modelo más complejo. A nuestro hijo – y a nosotros también – ahora le interesa cuando por fin tiene el álbum lleno y todos los espacios en blanco están rellenados con una foto. Dejamos de lado por el momento que los niños pueden intercambiar imágenes y así minimizar específicamente los espacios en blanco y utilizar la entrada total de manera muy eficiente. Definimos la duración por el número de paquetes necesarios. "Por casualidad", no podemos pensar en una solución algorítmica exacta para dicha probabilidad. Así que escribimos una breve simulación – igual para 10.000 niños diferentes – para obtener datos robustos. Tomamos valores reales de un álbum de cromos con k = 294 tarjetas adhesivas diferentes con futbolistas, escudos de clubes o trofeos, etc., etc. Los distintos tipos de cromos tienen el mismo estatus porque su adquisición es idéntico. También suponemos que cada imagen tiene la misma probabilidad de ser sacado, sino que cada paquete contiene siempre imágenes diferentes. Empecemos con la creación del conjunto de datos. El programa R `BL.sim()` saca cada caso de acuerdo con la las condiciones indicadas anteriormente hasta que se llene un álbum y, a continuación, devuelve los datos generados. El siguiente bucle simple llama a `BL.sim()` y genera para los 10 000 casos previstos simulaciones (`ptII_quan_classicstats_N-P_example-soccer-sim.r`).

```
seed <- 9876
set.seed(seed)
anzpp <- 5
Nkarten <- 294
trials <- 1e+4 #1e+5
maxIDs <- 1000
res.karten <- matrix(data=NA, nrow=trials, ncol=Nkarten)
res.IDs <- matrix(data=0, nrow=trials, ncol=maxIDs)
res.zaehl <- rep(0, trials)
for(i in 1:trials)
{
# print(i)
res <- BL.sim(Nkarten=Nkarten, anzpp=anzpp, doIDs=TRUE, total=FALSE) res.karten[i,] <-
res$karten
res.zaehl[i] <- res$zaehl
res.IDs[i,] <- res$IDs
}
```

En primer lugar, veamos el histograma y la estimación de densidad de la Figura 4.16. El eje X indica el número de compras hasta que un álbum está lleno. El eje Y contiene la correspondiente estimación de la

densidad de toda la distribución. La distribución es obviamente sesgada: los valores pequeños son más frecuentes que los muy grandes. En este caso, valdría la pena una transformación `log()` o `sqrt()` para que los datos fueran más simétricos (véase el capítulo 5.5.1 para un estudio de caso empírico), con lo que log() deja una impresión más adecuada (véase la Fig. 4.16).

```
# number of trials (=number of packages à five cards) till album is full str(res.zaehl)
# multiple cards - how many of each card are there at the end
# number of cards (double etc) for each card till zero
dim(res.karten)
# number of cards "to go"
# how many still required till album is full
dim(res.IDs) #rows=trials, cols=IDs
# IDs - distribution of necessary cards
# till album is full and nothing is missing
# extract number of necessary drawings for each trial/ case
laengeIDs <- apply(res.IDs,1, function(i) length(which(i != 0))) par(oma=c(2,1,1,1),
"cex.axis"=1, bty="l", mfrow=c(1,2)) hist(laengeIDs, panel.first=grid(),
           prob=TRUE, xlab="number of cards",
           ylab="Density", main="", col="skyblue", border="white")
lines(density(laengeIDs), col="red", lty=1, lwd=2)
# log() transfo of data
hist(log(laengeIDs), panel.first=grid(),
prob=TRUE, xlab="log(number of cards)",
           ylab="Density", main="", col="skyblue", border="white")
lines(density(log(laengeIDs)), col="red", lty=1, lwd=2)
mtext("Simulation soccer collector cards", outer=TRUE,
           line=-2, cex=1.5, side=3)
```

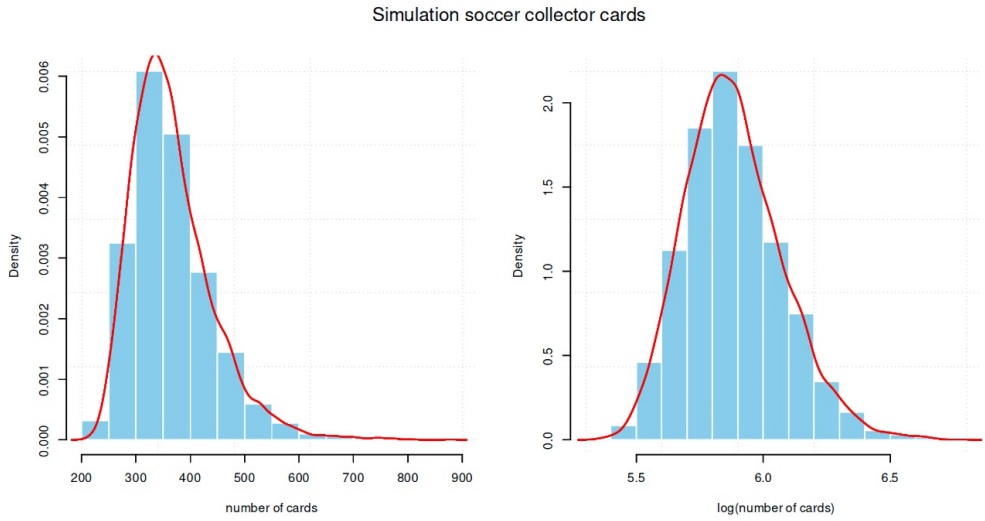


Figura 4.16. *Simulación – Recogida de imágenes de fútbol (histograma y estimación de la densidad)*

Ahora vale la pena mirar las cifras para ver cuántos paquetes se necesitan para completar el álbum de recortes.

```
> # number of packages (5 cards within each package)
> summary(res.zaehl)
Min. 1st Qu. Median Mean 3rd Qu. Max.
212  315     355    367  406     880
```

Según el resultado anterior, se necesitan al menos 199 paquetes para llenar el álbum. La media es 366,8 (mediana 355) y la máxima 866. Eso es bastante, suponiendo en el peor de los casos que tengamos que invertir 10,- Euros de compra por paquete. Sería de 1990,- a 8660,- Euros, con la mediana en 3550,- Euros y la media en 3668,- Euros. Así, el 50% de los compradores necesitarán 355 paquetes para tener todas las

tarjetas adhesivas. Un álbum de fútbol como éste es un asunto caro. Si lo convertimos en paquetes de compra a 0,70 euros, resulta el mínimo de 148,- Euros:

```
> # compare to .70-€ buying (each package)
> summary(res.zaehl)*.7
Min. 1st Qu. Median Mean 3rd Qu. Max.
148  220     248    257  284     616
```

Sigue siendo caro, pero significativamente menos que el cociente compra/paquete de 10,- Euros. Es decir, 148/(199 * 10) = 7:44%. Por término medio, llegamos a casi 257,- Euros – absolutamente considerado. Ahora bien, en principio puede reducirse, ya que en la práctica suelen distribuirse más tarjetas por compra de lo que indica la publicidad. Además, a veces hay otros adultos amigos que entregan a los padres con hijos sus propias tarjetas. Ahora bien, los intervalos de confianza son interesantes, ya que, en primer lugar, los cuantiles de los simulados de distribución. Los cuantiles son adecuados en este caso porque la distribución está sesgada (véase más arriba, Figura 4.16):

```
> # quantiles
> probs <- c(0.01,0.025,0.05,c(seq(0.1,1,by=0.1)),0.95,0.975,0.99)
> quantile(res.zaehl, probs=probs)
1%  2.5% 5%  10% 20% 30% 40% 50% 60% 70% 80% 90% 100% 95% 97.5% 99%
247 258  270 285 306 323 338 355 371 392 420 466 880  507 551   605
> sd(res.zaehl)
[1] 75.4
```

¿Cuánto tiempo necesitamos para todos estos paquetes? Los cromos no existirán para siempre. Si, por ejemplo, nuestra muestra compra por un precio ficticio de 85,- Euros (= mediana supuesta) a la semana en la tienda de descuento correspondiente (comestibles, productos de primera necesidad, etc.), el resultado para los distintos cuantiles es de semanas:

```
round(quantile(res.zaehl, probs=probs)*10/85)
```

La duración oscila entre 29 y 70 semanas, con una mediana de 42 semanas. Es decir, al menos medio año (29 * 7 / 365 = 0,5 años) y una mediana de 0,8 años, es decir, más de 9 meses:

```
round(quantile(res.zaehl, probs=probs)*10/85*7/365,2)
```

¿Cuántas tarjetas duplicadas hay? Para ello podemos fijarnos en cuántos múltiplos hay por diferentes tarjetas. La figura 4.17 muestra el número medio de tarjetas múltiples (dobles, triples, etc.) por número de tarjeta:

```
res.karten.ratio <- apply(res.karten,2,sum)/trials R-Code
summary(res.karten.ratio)
sd(res.karten.ratio)
plot(res.karten.ratio, panel.first=grid(), main="Multiple cards",
           xlab="card number",ylab="average number of multiples",col="red",
           type="h", bty="l")
```

Se puede observar que los cromos múltiples en todos los ensayos varían entre 6.17 y 6.32 (s = 0.025), que en última instancia refleja la base de la distribución equitativa de extracciones por carta (véase la Fig. 4.17), por lo que cabe suponer que es eficaz. Comparable sería un `heatmap()` de los datos brutos (no mostrados), que debería mostrar un patrón aleatorio:

```
heatmap(res.karten, Rowv=NA, Colv=NA, labRow=NA, labCol=NA, R-Code
xlab="card number", ylab="trials", main="heatmap")
```

En detalle, hay tarjetas únicas o tarjetas que están disponibles hasta 27 veces (máximo de varias tarjetas):

```
# double/ triple/ etc. complete
res.karten.tab <- table(res.karten)
```

```
res.karten.tab
res.karten.tab/trials
```

y el valor medio relacionado con los 294 cromos diferentes

```
> # mean of multiple cards
> sum(res.karten)/trials/Nkarten
[1] 6.25
```

La tarea complementaria consiste en examinar cuántas tarjetas nuevas y no redundantes cabe esperar en total por intento, es decir, hasta que un niño haya llenado el álbum. Esta corresponde (invertida) a la curva que muestra cómo el número de cartas que faltan disminuye con cada nueva sobre (véase la Fig. 4.18).

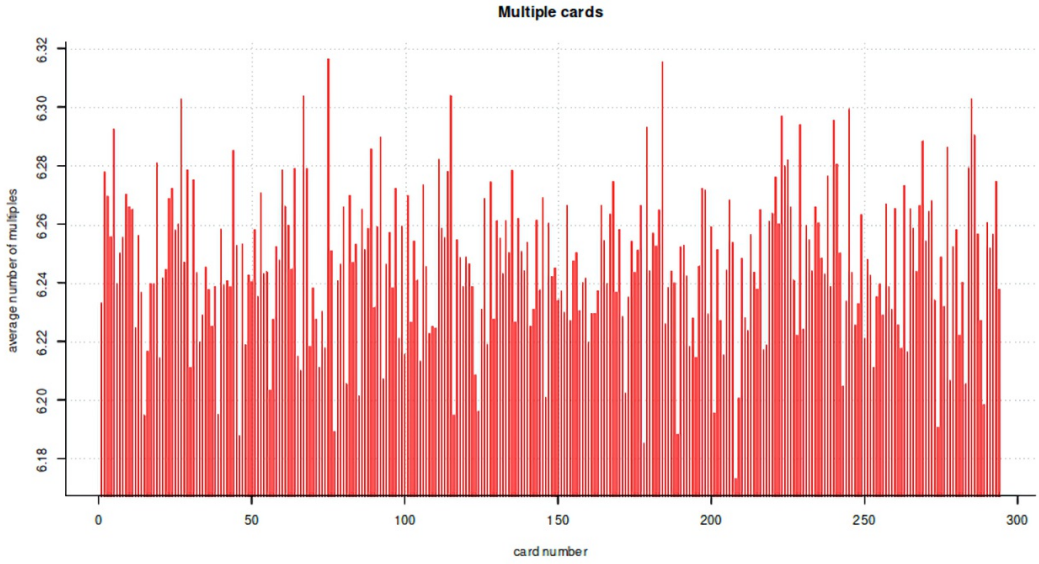


Figura 4.17. *Simulación – Recopilación de imágenes de fútbol (tarjetas múltiples)*

```
# new cards
res.IDs.sum <- apply(res.IDs,2,function(x) sum(x[!is.na(x)]))
res.IDs.mean <- apply(res.IDs,2,function(x) mean(x[!is.na(x)]))
res.IDs.max <- apply(res.IDs,2,function(x) max(x[!is.na(x)]))
res.IDs.min <- apply(res.IDs,2,function(x) min(x[!is.na(x)]))
res.IDs.sd <- apply(res.IDs,2,function(x) sd(x[!is.na(x)]))
# plot
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l")
plot(res.IDs.mean[is.finite(res.IDs.mean)], panel.first=grid(),
           main="", xlab="drawing",ylab="missing cards",
           col="red", type="l", bty="l")
lines(res.IDs.max[res.IDs.max != Inf], col="blue")
lines(res.IDs.min[res.IDs.min != Inf], col="darkgreen")
abline(h=29,col="magenta", lty=2)
abline(h=10,col="brown", lty=3)
legend("topright", legend=c("missing cards (mean)",
           "missing cards (max)","missing cards (min)",
           "n=29 cards-to-go", "n=10 cards-to-go"),
           pch="---", pt.cex=2.5,
           col=c("red","blue","darkgreen","magenta","brown"),
           bty="n", bg="yellow")
mtext("Process cards", outer=TRUE, line=-2, cex=1.5, side=3)
```

Además, la figura 4.18 contiene dos líneas horizontales, a saber, cuánto tiempo (operacionalizado como paquetes perdidos) sigue faltando cuando n = 29 (= 10%) o n = 10 (= 3:4%) de las tarjetas siguen desaparecidos. En ambos casos, la progresión cuasi-asintótica hacia el final conduce a inversiones despro-

porcionadamente grandes inversiones hacia el final, que deben acortarse. El guión R contiene otras posibilidades de representación.

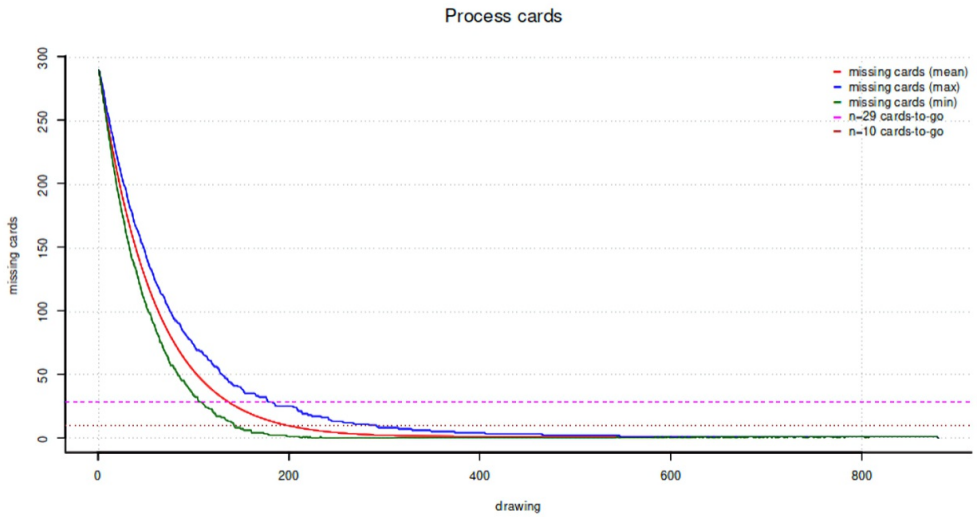


Figura 4.18. *Simulación – Colecciona fotos de fútbol (mapa de progresión al álbum completo)*

Del mismo modo, cabe preguntarse cómo se recorre el camino desde el punto de partida de un álbum vacío hasta un álbum completo. Para simplificar, el análisis se limita a un único conjunto de datos (véase la Fig. 4.19). Es importante señalar que, al reproducir estos conjuntos de datos, puede darse el caso de que en la versión de R o los paquetes de R utilizados los generadores aleatorios incorporados pueden producir valores diferentes con el mismo valor inicial, por lo que las cifras impresas aquí pueden no poder reproducirse exactamente o incluso desviarse considerablemente. Por lo tanto, el Apéndice B.3 contiene información sobre la versión de R, la información de la sesión y el sistema operativo utilizado.

```
# single R-Code
# IDs
seed <- 9876
set.seed(seed)
res.single <- BL.sim(Nkarten=Nkarten, anzpp=anzpp, doIDs=TRUE, total=TRUE)
str(res.single)
waytoZ <- res.single$IDs
waytoZ <- waytoZ[!is.na(waytoZ)]
plot(waytoZ, panel.first=grid(), type="l", col="red", main="",
           xlab="Simulation (packages)", ylab="number of missing cards", bty="l")
mtext("Card collection process", outer=TRUE, line=-2, cex=1.5, side=3)
```

Aquí los números para el caso 29 tarjetas todavía faltantes:

```
> # 29 left
> abline(h=29,col="blue", lty=2)
> # 10 left
> abline(h=10,col="darkblue", lty=3)
> # duration time (in packages) for 29 and 10 missing cards left
> # which means
> Nkarten - 29
[1] 265
> Nkarten - 10
[1] 284
> # are already present
> head(waytoZ)
[1] 289 284 279 275 271 266
```

```
> tail(waytoZ)
[1] 3 3 3 1 1 0
> waytoZ.l <- length(waytoZ)
> waytoZ.l
[1] 235
> which(waytoZ == 29)
[1] 130 131
> waytoZ.l - which(waytoZ == 29)
[1] 105 104
> which(waytoZ == 10)
[1] 191 192
> waytoZ.l - which(waytoZ == 10)
[1] 44 43
```

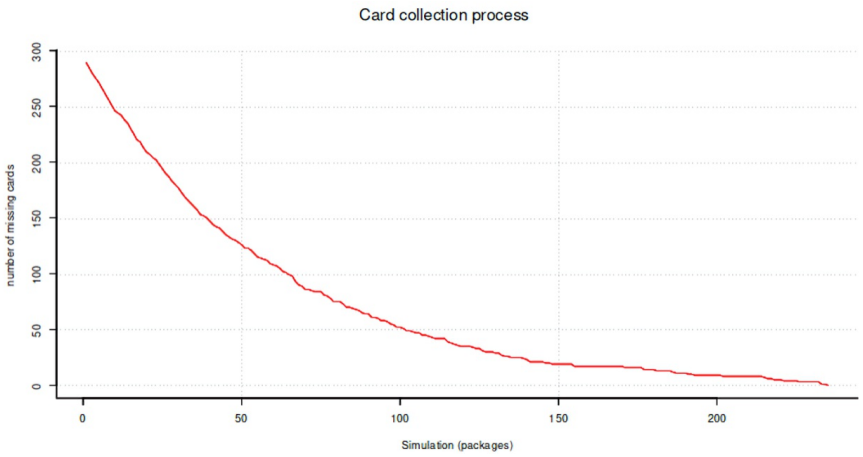


Figura 4.19. *Simulación – Coleccionar cromos (mapa de progresión a álbum completo – caso único)*

Se puede observar que, por ejemplo, faltando aún 29 cartas (es decir, 294 - 29 = 265 cartas presentes), ya hay 131 sobres presentes (= 131 * 5 = 655 cartas en total), pero aún se necesitan 235 - 131 = 104 paquetes hasta que el álbum esté lleno, es decir, con (129/294) *100% = 90.1% de cartas presentes, sólo ha transcurrido 131/235 *100% = 55.7% del tiempo hasta que el álbum esté lleno – medido en la duración determinada empíricamente por simulación. Aquí, en el ejemplo de simulación única, la duración total es de 235 paquetes. Sin embargo, esto no es la regla, sino un caso individual y puede llevar menos tiempo, pero también bastante más. De forma equivalente, puede calcularse para 10 tarjetas que falten (véanse los datos brutos anteriores en la R) o para cualquier número de tarjetas que falten.

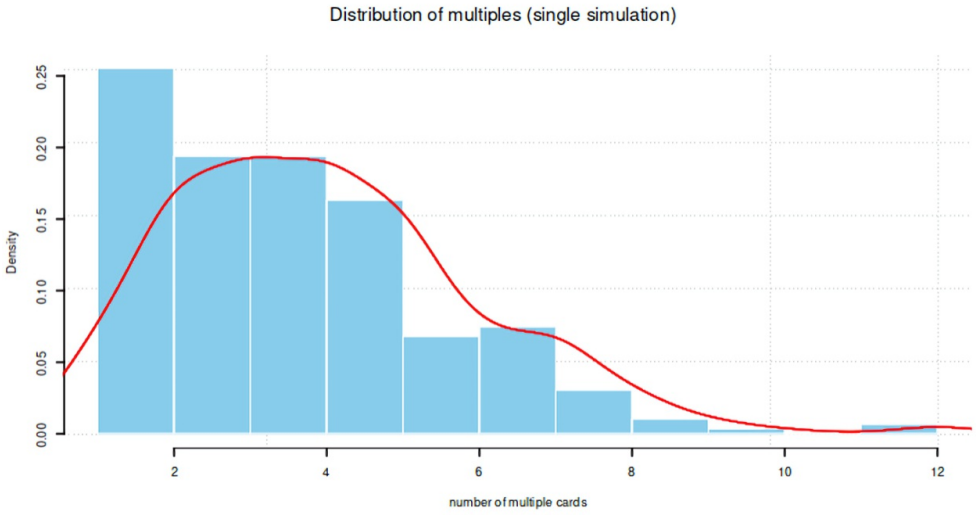


Figura 4.20. *Simulación – Coleccionar cromos – caso único*

```
# distribution of each card R-Code
karten.total <- res.single$karten.total
karten.total <- karten.total[!is.na(karten.total)]
plot(karten.total, panel.first=grid(), type="p", bty="n")
kt.tab <- table(karten.total)
table(kt.tab)
kt.tab.fn <- fivenum(kt.tab)
names(kt.tab.fn) <- c("Min","LQ","Median","UQ","Max")
kt.tab.fn
kt.tab
hist(as.vector(kt.tab), panel.first=grid(), prob=TRUE,
           col="skyblue", border="white",
           main="", xlab="number of multiple cards", ylab="Density")
lines(density(as.vector(kt.tab)), col="red", lwd=2)
mtext("Distribution of multiples (single simulation)",
           outer=TRUE, line=-2, cex=1.5, side=3)
```

¿Cómo puede mejorarse fundamentalmente el modelo en términos de predicción? Es más difícil en términos estadísticos clásicos porque *no se nos permite incluir el conocimiento previo*. Por lo tanto, tomamos el enfoque bayesiano como modelo de pensamiento y consideramos en primer lugar, en términos de contenido, qué posibilidades se nos abren para utilizar el conocimiento previo de forma significativa. A continuación, no nos centraremos en los análisis estadísticos, sino en el pensamiento cualitativo, qué factores pueden influir. Ya se han hecho algunas observaciones anteriormente, como el hecho de recibir más tarjetas por compra (por ejemplo, a través de dependientes amables, otros adultos). Naturalmente, surgen cuestiones como la distribución, la venta, el reparto de tarjetas en función de la duración de la acción y la inteligencia social. Por supuesto, se trata en gran medida de variables ficticias, pero podrían utilizarse de forma realista si se dispusiera de datos. Ninguno de estos datos está disponible aquí, por lo que sigue siendo un experimento mental de gran utilidad.

- *Distribución* – Descubrimos que los envases no se distribuyen por igual a todos los mercados. Algunos mercados tienen especialmente parcelas con un número de imágenes bajo, medio o alto. Esto es ficticio, por supuesto, pero podría influir en la distribución de probabilidad de los números en relación con las tiendas de la empresa de descuento, de modo que se podrían realizar compras selectivas en las tiendas para obtener determinados rangos de números con mayor probabilidad. Sin embargo, no hay garantía de ello. Con datos de fondo suficientes, se podría reconstruir una distribución de los rangos de números en las tiendas. Asimismo, si existe un cupo total limitado de tarjetas, se podría investigar qué número de las tarjetas disponibles con qué tirada se han entregado ya en cada caso y cuántas cajetillas se han distribuido por tienda. El resultado es una distribución individual por tarjeta y por sucursal o zona. A partir de estos datos, podría crearse a su vez una distribución compleja sobre cuántas tarjetas se distribuyeron en absoluto y qué rangos de números están especialmente representados dónde o no. Por consiguiente, las visitas a las sucursales tendrían que coordinarse con precisión, aunque sin la garantía de recibir exactamente las tarjetas que tienen más probabilidades de estar allí según la distribución. La compra concreta sólo supone una mejora relativa de las probabilidades y debe ajustarse constantemente en coordinación con las cartas ya disponibles.
- *Ventas* – Resulta que ciertos cajeros dan más y nos ceñimos a ellos en consecuencia. Esto acorta el tiempo o los costes de inversión. Lo mismo se aplica a estar atento a otros adultos que puedan no necesitar sus tarjetas, devolverlas o no quieren aceptarlas o incluso tirarlas. El argumento en contra de este planteamiento es no ser intrusivo.
- *Distribución y duración* – Es evidente que se reparten más tarjetas hacia el final de la campaña que al principio. Sin embargo, esto también debe entenderse en términos relativos. En ese caso, habría que hacer más compras en ese momento, aunque es difícil. Al fin y al cabo, los sobres se dan además de las compras normales, que no aumentan sólo porque pronto no habrá más cartas.
- *Inteligencia social* – Este es el componente más prometedor para conseguir llenar el álbum con el mínimo esfuerzo. Los niños aprenden a comerciar de forma justa. Como no se espera que a todo el mundo le falten las mismas tarjetas, hay muchas posibilidades de conseguir las fotos que faltan más rápidamente que a través de la compra, que carece por completo del aspecto cooperativo. Esto

requiere perseverancia, cierta inversión de tiempo y habilidades sociales para ser un socio de intercambio fiable. A partir de la prueba práctica, puede decirse que el trueque aportó efectivamente la máxima aceleración para conseguir las últimas cartas. Solo con la ayuda de las compras, los álbumes no estarían llenos y, de todos modos, la acción limitada en el tiempo termina mucho antes.

Todos estos factores podrían modelizarse numéricamente para crear un modelo complejo de distribución probabilística – por ejemplo, en relación con una zona o incluso un mercado concreto, cursos individuales y ajustados, etc. – que, en principio, podría utilizarse como base para un modelo de distribución principialista. En principio, podría crearse una distribución a priori (véase el capítulo 6.14.6) a partir de ella, que se utilizaría como conocimiento previo en los cálculos. El esfuerzo sería enorme y difícilmente justificable desde el punto de vista de un análisis coste-beneficio. Por tanto, la opción más prometedora parece ser el componente social, es decir, el intercambio de imágenes. Es fácil, no cuesta nada y es absolutamente eficaz, ya que sólo se intercambian las fotos que faltan y puede traer nuevos amigos. Y amplía el aspecto numérico para incluir el pensamiento cualitativo y la unión social, es decir, la *cooperación*.

Por último, pasamos a un caso real de la práctica. Aquí, el intercambio comenzó a partir de n = 29 tarjetas (véase la Fig. 4.18 para el tiempo total previsto en ese punto). Sólo intercambiando se llenó el álbum antes de que la tienda de descuento pusiera fin regularmente a la promoción. Como era de esperar, los últimos paquetes sólo contenían múltiplos y no más nuevas tarjetas. Después de 294 (= necesarias) + 456 (= redundantes) tarjetas, el álbum estaba lleno, de modo que se acumuló un total de 750 imágenes, correspondientes a 150 paquetes.

4.3.6 Tamaño de la muestra

Las diferencias o correlaciones relevantes deben tener cierto nivel de significación, apuntar en una dirección clara y estar relacionadas con el contexto. En términos de estadística frecuentista, el tamaño de la muestra debe ser proporcional, como en el capítulo 4.4.13 muestra lo que ocurre cuando una muestra simplemente se hace grande y se infla. Si hay un efecto mínimo, ya sea en la dirección "correcta" o no, todo se vuelve altamente significativo desde el punto de vista estadístico si N es lo suficientemente grande. Una diferencia arbitrariamente pequeña se encontrará con una probabilidad de 1 en una muestra arbitrariamente grande, y eso es trivial. El tamaño del efecto permanece – pero constante dentro de la variación de muestreo habitual (véase Fig. 4.66, p.306) y puede resultar completamente insignificante en la escala original. En cambio, los tamaños del efecto no son sensibles a las variables independientes ni a su variación prevista (es decir, la "manipulación"). Medir el grado de cambio producido de este modo aún todavía no es fácilmente posible (Ronis, 1981; Brandstätter, 1999, p.13).

Por el contrario, si una muestra es muy pequeña, no puede calcularse razonablemente en la estadística clásica, o sólo de forma limitada. Esto no se aplica a la estadística bayesiana, que llega a conclusiones coherentes incluso con un N pequeño (Studer, 1996b), aunque la incertidumbre sea entonces naturalmente mucho mayor que con un N suficiente. El arsenal disponible de análisis estadísticos se reduce, ya que suele exigirse un tamaño mínimo de muestra para poder utilizar un procedimiento. Además, la hipótesis nula se favorece a sí misma de forma poco realista al elegir límites de significación pequeños. Además, las hipótesis nulas empíricamente teóricamente significativas y verdaderas son prácticamente desconocidas, por ejemplo, en la investigación social empírica y en muchas otras disciplinas. Gelman (2011b) lo resume diciendo que

„[...] is that the hypothesis of zero effect is almost never true! The problem with the significance testing framework – Bayesian or otherwise – is in the obsession with the possibility of an exact zero effect. The real concern is not with zero, it's with claiming a positive effect when the true effect is negative, or claiming a large effect when the true effect is small, or claiming a precise estimate of an effect when the true effect is highly variable, or ..."

Encontrar efectos inexistentes es importante en el sector farmacéutico. Por ejemplo, se necesitan pruebas para demostrar la ausencia de efectos secundarios en comparaciones de grupos experimentales (por ejemplo,

nuevo fármaco frente a placebo o fármaco existente). Pero una ganancia teórica de conocimiento no surge realmente del no rechazo de la hipótesis nula. Peor aún: puede ocurrir que una muestra aleatoria produzca valores extremos y, a pesar del tamaño mínimo de la muestra, surjan valores p pequeños como propiedades de los datos, sobrestimando un efecto existente pero mínimo y prácticamente insignificante. Además, existe el riesgo de que se produzcan errores de tipo S y M (véase el capítulo 4.3.3.2). En este caso, como en muchas otras ambigüedades, sólo la repetición y una definición precisa de las características de la muestra pueden ayudar – así como un cuidadoso diseño y formulación de hipótesis.

Para encontrar diferencias en grupos pequeños, hay que fijarse muy bien. Hay que cambiar la lupa por el láser para ver las sutiles diferencias. Sin embargo, encontrar diferencias cuanto más de cerca se mira suele ser trivial, como muestra el debate sobre el tamaño del efecto. Si se encuentran diferencias, siempre se debe preguntar inmediatamente por la dirección y la magnitud (errores de tipo S y de tipo M, Gelman & Carlin, 2014, véase también Gelman & Weakliem, 2009).

Cualquier resultado puede ser un raro acontecimiento aleatorio, porque no lo sabemos. Este es el precio de la estadística clásica: *todo es azar*, ya que no se permite por definición una dirección u orientación preestablecida en forma de conocimiento previo. Esto significa que un instrumento de medición debe ser muy preciso, es decir, debe tener una potencia de prueba *muy elevada*. Si un método de medición es sensible a las diferencias, puede utilizarse con una muestra más pequeña, siempre que se replique. Si no es así, se necesita una muestra mayor para encontrar el mismo efecto. Si el efecto en la población es grande, basta con una muestra más pequeña y un procedimiento de medición menos sensible. Pero todo esto no lo sabemos de antemano. La relación concreta de las interdependencias descritas se determina mediante un análisis de poder a priori según Neyman-Pearson. En este proceso, las tasas de error α y β suelen determinarse de antemano y se define el tamaño del efecto deseado *d*. A partir del conocimiento del procedimiento de análisis estadístico de datos que se va a utilizar, hay que derivar un tamaño de muestra adecuado, lo que a su vez requiere una definición precisa del contenido de esta muestra. Un ejemplo clásico sería la representatividad de una muestra en el contexto de unas elecciones basada en el conocimiento de la composición de la población votante y las variables efectivas que influyen en el comportamiento electoral (sexo, edad, ocupación, nivel de educación, ingresos, lugar de residencia y región, etc.). A continuación se procede a la aplicación empírica.

La estadística de Bayes permite trabajar con muestras muy pequeñas y no sólo permite, sino que esencialmente exige la inclusión del conocimiento contextual como información previa. Esto nos lleva a la clásica acusación de subjetividad y, en el reverso de esta acusación, a la pregunta de si la teoría Neyman-Pearson y la estadística clásica en general es en realidad una teoría objetiva. Antes de entrar en detalles, veamos un ejemplo de la influencia del tamaño de la muestra.

En primer lugar, modelamos con código R las dependencias de los valores *t*, los errores estándar, el tamaño de la muestra y los datos del tamaño del efecto, ya que estos últimos no se dejan impresionar por el aumento del tamaño de la muestra. Se extraen muestras con $\mu = 100$ y $\sigma = 10$ de una población conocida con distribución normal. Se extraen repetidamente $n = 100$ muestras aleatorias de esta población y se realiza una prueba *t* de esta muestra aleatoria tanto frente a cero (prueba de significación convencional) como contra $\mu = 100$. En cada repetición, el tamaño de muestra *N* se incrementa en $n = 10$, de modo que esta secuencia contiene tamaños de muestra de 5 a 995. Además, el código genera por simulación el tamaño del efecto *d* de Cohen (`ptII_quan_classicstats_N-P_SE-N-dep.r`).

```
N <- seq(5,1000,10)
N.l <- length(N)
set.seed(1234)
mu1 <- 100
sigma1 <- 10
ses <- do.call("rbind", lapply(seq_along(N), function(i)
{
samp <- rnorm(n=N[i], mean=mu1, sd=sigma1)
mw <- mean(samp)
sabw <- sd(samp)
se <- sabw/N[i]
tv <- t.test(samp, alternative=("two.sided"))[["statistic"]]
tv.pop.mw <- t.test(samp, alternative=("two.sided"),
           mu=mu1)[["statistic"]]
```

```
d <- (mw-mu1)/sigma1
return(c(mu1,mw,sigma1,sabw,se,N[i],tv,tv.pop.mw,d))
}))
colnames(ses) <- c("mu","mean","sigma","sd","se","N","t","t.mu","d")
```

Los datos brutos tienen este aspecto

```
head(ses) R-Code
tail(ses)
```

y puede describirse con las estadísticas habituales del siguiente modo:

```
describe(ses)
```

Relevantes y significativos son los diferentes gráficos (véase la Fig. 4.21) de las estadísticas generadas entre sí:

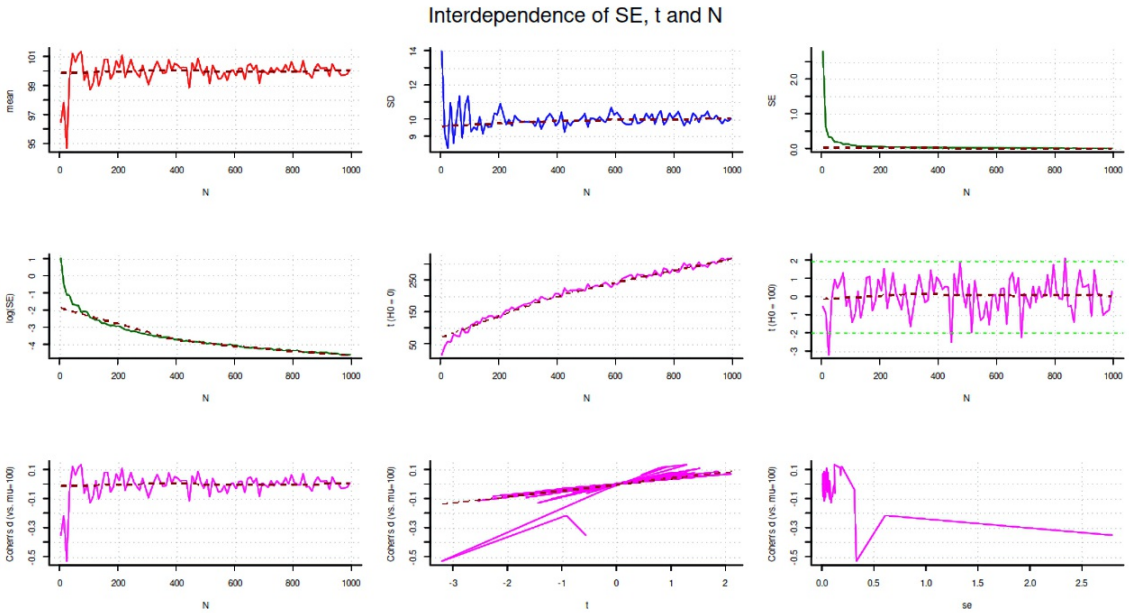


Figura 4.21. *Relación entre los valores t, errores estándar SE, d de Cohen y tamaño de muestra N*

```
# par(mar=c(5,6,5,5))
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l", mfrow=c(3,3))
plot(N,ses[,"mean"], panel.first=grid(), type="l",
           col="red", ylab="mean")
lines(lowess(ses[,"mean"] ~ N), lty=2, lwd=1.4, col="darkred")
plot(N,ses[,"sd"], panel.first=grid(), type="l",
           col="blue", ylab="SD")
lines(lowess(ses[,"sd"] ~ N), lty=2, lwd=1.4, col="darkred")
plot(N,ses[,"se"], panel.first=grid(), type="l",
col="darkgreen", ylab="SE")
lines(lowess(ses[,"se"] ~ N), lty=2, lwd=1.4, col="darkred")
plot(N,log(ses[,"se"]), panel.first=grid(), type="l",
           col="darkgreen", ylab="log(SE)")
lines(lowess(log(ses[,"se"]) ~ N), lty=2, lwd=1.4, col="darkred")
plot(N,ses[,"t"], panel.first=grid(), type="l",
           col="magenta", ylab="t (H0 = 0)")
lines(lowess(ses[,"t"] ~ N), lty=2, lwd=1.4, col="darkred")
plot(N,ses[,"t.mu"], panel.first=grid(), type="l",
           col="magenta", ylab=paste("t (H0 = ",mu1,")",sep=""))
lines(lowess(ses[,"t.mu"] ~ N), lty=2, lwd=1.4, col="darkred")
abline(h=c(-1.96,1.96), col="green", lty=2, lwd=0.8)
```

```
plot(N,ses[,"d"], panel.first=grid(), type="l",
           col="magenta", ylab=paste("Cohen's d (vs. mu=",mu1,")",sep=""))
lines(lowess(ses[,"d"] ~ N), lty=2, lwd=1.4, col="darkred")
plot(ses[,"t.mu"],ses[,"d"], panel.first=grid(), type="l", col="magenta",
xlab="t", ylab=paste("Cohen's d (vs. mu=",mu1,")",sep=""))
lines(lowess(ses[,"d"] ~ ses[,"t.mu"]), lty=2, lwd=1.4, col="darkred")
plot(ses[,"se"],ses[,"d"], panel.first=grid(), type="l", col="magenta",
           xlab="se", ylab=paste("Cohen's d (vs. mu=",mu1,")",sep=""))
lines(lowess(ses[,"d"] ~ ses[,"SE"]), lty=2, lwd=1.4, col="darkred")
mtext("Interdependence of SE, t and N", outer=TRUE, line=-2, cex=1.5, side=3)
```

En la Figura 4.21 se puede observar que para los conjuntos de datos simulados con un tamaño de muestra creciente

- los valores medios se nivelan y tienden hacia el valor poblacional μ
- las desviaciones típicas también se nivelan y tienden al valor poblacional σ
- los errores estándar tienden a cero, por lo que vale la pena trazar el mismo gráfico en la escala log().
- los valores *t* crecen constantemente cuando se comparan con cero
- los valores *t* de la prueba frente al valor poblacional oscilan en torno a cero dentro de un rango manejable
- las *d* de Cohen (potencias de efecto) oscilan en torno a cero en un marco muy manejable
- la *d* de Cohen (potencias de efecto) frente a los valores *t* (prueba contra μ) se agrupan en torno a cero
- la *d* de Cohen (potencias de efecto) frente a los errores estándar se agrupan en torno a cero.

De ello se deduce, según los gráficos de la figura 4.21, que a medida que aumenta *N*, cada vez que se encuentra una pequeña desviación de Cero, los errores estándar se reducen hacia cero o los valores *t* crecen hacia el infinito cuando se realizan pruebas frente a Cero. Con la prueba *t* frente al parámetro poblacional μ (que a menudo se desconoce en la investigación en ciencias sociales o psicología), los valores se comportan como la potencia de efecto: oscilan en torno al parámetro de población, que corresponde a un efecto de cero. Las potencias de efecto, por su parte, permanecen relativamente constantes, ya que también se aproximan a su valor poblacional, que en este caso también es Cero. Sin embargo, queda claro que el tamaño de la muestra no influye realmente en la d de Cohen, sino que el efecto del tamaño de la muestra en el parámetro de población es más significativo. Más bien ocurre que las d oscilan en torno al parámetro poblacional Cero.

A nivel numérico, la lógica de la estadística clásica es evidente. El número de valores *t* superiores al criterio habitual del 5% ($t \approx 1{,}96$) difiere para la prueba contra el μ

```
> #how many t values are greater than critical value?
> alpha <- 0.05
> sum(abs(ses[,"t.mu"]) > qnorm(1-alpha/2))/N.l
[1] 0.05
```

o contra Cero

```
> sum(abs(ses[,"t"]) > qnorm(1-alpha/2))/N.l
[1] 1
```

En la prueba contra μ, como era de esperar, alrededor del 5% se desvía de esto. En la prueba contra Cero, casi todos los valores *t* se desvían de Cero de forma estadísticamente significativa. Por término medio, las medias empíricas y las desviaciones estándar ya se aproximan bastante a los valores de la población μ y σ:

```
> media(ses[, "media"])/mu1
[1] 0.999
> media(ses[, "sd"])/sigma1
[1] 0.999
```

Es importante añadir que en una comparación de dos grupos, si los dos grupos difieren realmente, los valores numéricos serán diferentes. Entonces, las potencias de efecto no oscilarán en torno a Cero, sino en torno a la diferencia realmente presente. El presente ejemplo sólo incluye el caso de un grupo. Sería una tarea para los lectores cambiar el código R de tal manera, que se producen tamaños de muestra aún mayores y, a continuación, ampliar el código R para producir una simulación para el caso de dos grupos. A continuación, se comprobará la validez de las afirmaciones anteriores a partir de los gráficos generados.

Caso 4.4: Problemas de diseño

Imaginamos que dos universidades participen en un estudio de caso de investigación educativa empírica y entreguen juntos una solicitud para ganar financiación de terceros para costes de ca. 500'000.- Euros, una duración de tres años y varios puestos (media jornada y jornada completa). Los objetivos formales son varias publicaciones impor-tantes para la cualificación profesional de los autores. En cuanto al contenido, el objetivo es presentar métodos cooperativos de enseñanza y aprendizaje en el aula. El estudio consiste en la formación continua de los profesores, y los datos se refieren sobre todo a mediciones del rendimiento de los alumnos (Realschule, 7º curso). El tamaño de la muestra es bastante pequeño, ya que sólo una parte de las clases de las escuelas destinatarias pueden incluirse en el diseño y algunas de los temas a estudiar incluso tienen que cancelarse por completo. Además, se dispone de menos fechas de encuesta de lo que sería razonable.

Desgraciadamente, la propuesta no tiene en cuenta el hecho de que tales estudios deben evaluarse con HLM/MLM, sino que asume la repetición de pruebas t repetidas simples. Una suposición fatal, no sólo porque las repetidas pruebas t no son muy científico y por lo general sólo pueden contrarrestarse con un número creciente de correcciones de Bonferroni. Esas comparaciones masivas de medias no tienen en cuenta en absoluto la complejidad de la realidad, que sin duda puede modelizarse mediante las variables recogidas. Este error inicial tiene un efecto enorme más adelante. Los análisis realizados hubieran requerido en realidad una muestra más grande para poder modelar modelos HLM que cubren con el número de puntos de medición y el tamaño de las muestras disponibles. Por consiguiente, no existen las significaciones esperadas salvo unas pocas, que no se puede tomar en serio hasta cierto punto. Pueden identificarse potencias de efecto de tamaño suficiente y en una dirección que se ajuste a la hipótesis, pero debido a la falta de replicación y al pequeño tamaño de la muestra o al reducido número de puntos de medición, existe una gran incertidumbre sobre la estabilidad y la gravedad de los efectos. Al final, habría que preguntarse: ¿justifica el resultado tanto el uso de fondos de terceros o son una razón para utilizar ahora los métodos de enseñanza y aprendizaje antes mencionados a gran escala y con grandes gastos en la formación y perfeccionamiento del profesorado?

La pregunta crítica de la tarea 4.4 puede responderse con un "puede ser, puede no ser", una situación clásicamente insatisfactoria, porque no se puede dar una respuesta clara. Al mismo tiempo tiene poco que ver con la calidad de las intervenciones o el diseño del estudio. El fallo radica claramente en la metodología-estadística y la cuestión de si la cantidad de datos permite formular hipótesis de forma más clara e inequívoca posible, con un mínimo de incertidumbre. En este tipo de estudios los cálculos de los tamaños de muestra correspondientes deben comprobarse varias veces y los análisis deben realizarse de acuerdo con los últimos estrategias analíticas.

Lo que ahora debería estar claro es que el aumento del tamaño de las muestras hace que se encuentren diferencias existentes con la máxima probabilidad. Sin embargo, si observamos el otro extremo de la distribución del tamaño de las muestras (véase la Fig. 4.21), veremos lo caóticos y extremos que pueden ser los valores cuando las muestras son pequeñas. La *d* de Cohen muestra entonces efectos putativos, los errores estándar son muy grandes y los valores medios se desvían significativamente de la media de la población. Dependiendo del contexto real y del tema, el muestreo aleatorio puede revelar desviaciones aún más significativas, que como tales no son fácilmente aparentes como desviaciones aleatorias. Por lo tanto – como también muestra el gráfico, cuando los valores medios cambian al máximo de una muestra a otra – siempre hay que ser prudente a la hora de generalizar conclusiones ante resultados de muestras pequeñas y no

replicadas. Por tanto, esto se aplica a "muchísimos estudios psicológicos" que tradicionalmente se basan en muestras pequeñas.

4.3.7 La subjetividad en la estadística clásica

La inclusión de información contextual a priori en las ecuaciones se rechaza por completo desde la perspectiva de la estadística clásica y se descarta y devalúa por *subjetiva*, aunque no se puede criticar la matemática subyacente (Jaynes, 2003, p.493).

> „In fact, they [Se refiere a estadísticos clásicos como Fisher, Kendall, Cramér, Feller o Neyman – nota de los autores.] offered no demonstrative arguments or factual evidence at all in support of their position; they mearely repeated ideological slogans about 'subjectivity' and 'objectivity' which were quite irrelevant to the issues of logical consistency and useful results."

En cambio, en la estadística clásica, los umbrales de significación y muchas otras cosas se fijan, en el mejor de los casos, por convención y casi nunca se basan en consideraciones sustantivas. Esto es irreflexivo, no se basa en el objeto de estudio y, desde luego, no es un ejemplo de enfoque objetivo. Para una mejor comprensión, añadimos aquí que un enfoque supuestamente objetivo es idealmente aquel en el que el ser humano no ejerce una influencia distorsionadora. En el presente caso, cabe señalar que las decisiones qua convención no cumplen este criterio de objetividad. Objetivo sería cuando la elección de los parámetros resulta o puede derivarse directamente del objeto de investigación. Esto, a su vez, está lejos de ser posible a pesar de los algoritmos de inteligencia artificial. Sin seres humanos, no se pueden tomar decisiones. La cuestión sería si hacemos uso de las capacidades humanas, que incluyen la subjetividad como característica humana central, o no. Pensar en ello y llenarlo de ejemplos concretos sería tarea de los lectores. No hay que olvidar que tanto los instrumentos de ensayo como los algoritmos de análisis, etc., son obra del hombre. La subjetividad, aunque esté oculta, siempre se abre paso en estos ámbitos.

Curiosamente, la estadística de Bayes no sólo se utiliza en las ciencias sociales, sino también en la física cuántica, la física del plasma y la astrofísica, es decir, en disciplinas que parecen estar bastante por encima del término "subjetividad". Sin la estadística de Bayes, estas disciplinas (por ejemplo, las probabilidades cuánticas o los sucesos singulares en astrofísica, Loredo, 1990) no podrían funcionar, ya que la estadística clásica no suele ser aplicable en estos casos, pues no puede trabajar con pocos casos o incluso con casos únicos. Además, el conocimiento contextual siempre fluye, de lo contrario no se podría planificar y llevar a cabo una investigación. La acusación inespecífica de subjetividad, operacionalizada concretamente como una distribución de probabilidad a priori de los parámetros sobre la base del conocimiento experto y la información contextual disponible, debe integrarse a este respecto en las cuestiones de

- Cribado de la información disponible sobre el objeto de investigación
- Justificación en las referencias teóricas
- Contenido de la información, es decir, expectativas sobre la influencia (magnitud y dirección).
- Transparencia de la aplicación
- Documentación para garantizar la repetibilidad
- Repetición, para sustituir el conocimiento contextual por datos empíricos dividirse con respecto a la forma en que se procesa la misma información contextual.

Además, lo objetivo es al mismo tiempo intersubjetivo, ya que nada es posible sin investigar temas, como la toma de decisiones, la programación de algoritmos, etc.

La subjetividad es relativa. Por lo tanto, como ya se ha explicado, existen argumentos razonables para acusar a la estadística clásica de tomar algunas decisiones, o incluso muchas, subjetivas; por ejemplo, la elección del nivel de significancia, que presumiblemente en más del 99,999 % de todos los estudios

publicados no estaba ni está nunca justificada de forma sustantiva, sino que se aplica sin mucha reflexión qua convention. Incluso las diferencias y sus consecuencias en los habituales 5%, 1%, 0,1% nunca se discuten en el contexto de la pregunta de investigación. Esto indica no sólo subjetividad, sino falta de orientación hacia los problemas. A la luz de las declaraciones tardías de Fisher (véase el capítulo 4.3.2), ya no se le pueden reprochar póstumamente las convenciones practicadas. No se puede criticar el hecho de que no sólo en el marco de las pruebas estadísticas clásicas deba crearse un criterio que distinga entre estadísticamente significativo e insignificante y cree así una base para la decisión sobre una prueba. Es más bien una cuestión de aplicación de esta necesidad. El mismo problema se aplica igualmente a la estadística de Bayes, y en ella encontramos desde hace tiempo un debate similar, tanto sobre la cuestión fundamental de la subjetividad como a nivel concreto, por ejemplo, sobre los factores de Bayes (véase el capítulo 6.8.1). Este debate plantea la cuestión de si las pruebas son la metodología adecuada para abordar los datos. Hay autores que sitúan la modelización de relaciones complejas por encima de las pruebas, lo que no significa que deban prescindir de ellas ante una pregunta de investigación que requiera pruebas. La complementariedad en el sentido de "ambos y" en lugar de "uno o" ofrece una alternativa a un ámbito más limitado en estadística. La elección de un procedimiento de análisis de datos y sus límites de parámetros significativos pertenece, pues, a la razón y depende directamente del problema. Lo mismo cabe decir de los criterios de corte elegidos para determinar la significación práctica y estadística. Esto es cierto independientemente de la teoría estadística elegida.

Este problema tan debatido puede localizarse directamente en el proceso de trabajo de la investigación en Gigerenzer (1981) y en una ampliación en Gürtler (2005). Se trata de la traslación del relativo empírico al relativo numérico y viceversa: del numérico al relativo empírico (véase la Fig. 4.22). Todos los números de las estadísticas adquieren su significado en estos dos pasos, cuando primero la operacionalización de la pregunta de investigación abre el espacio de posibilidades de los resultados y cuando después los resultados numéricos se vuelven a relacionar con las preguntas de fondo. Y para ello se necesitan criterios. Esto sólo puede hacerse mediante un examen sustantivo de la materia y nunca por convención. Además, a menudo sólo se entiende parcialmente por qué se utiliza un determinado método estadístico y no otro muy similar. Falta una justificación última, que no puede esperarse realmente en vista de un criterio relativo de verdad (véase el capítulo 1.1).

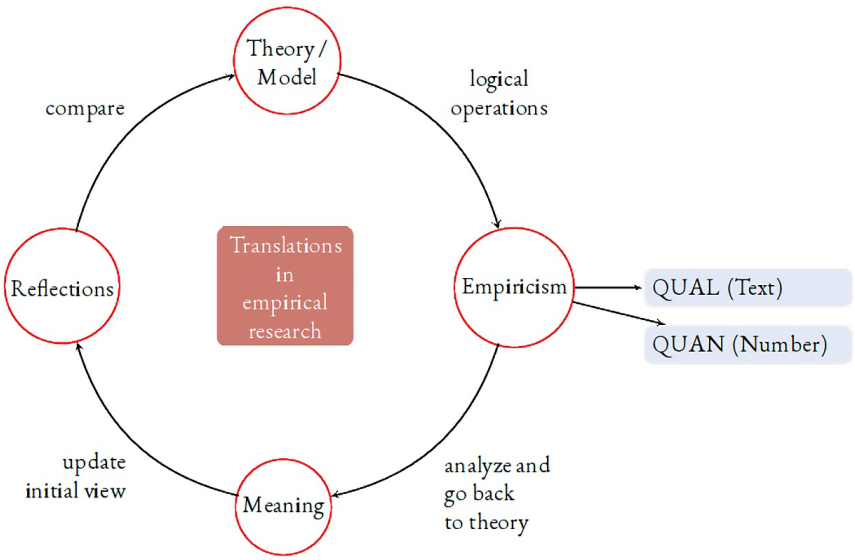


Figura 4.22. *Las traducciones en el proceso de investigación científica*

La subjetividad como característica no debe realmente causar dificultades – siempre que se explicite y documente las decisiones. La realidad en sí misma es relativa y no existe independientemente de nosotros en el sentido de que no podemos observarla e investigarla sin un observador. Si realmente existe y qué es,

lo consideramos poco interesante e irrelevante desde un punto de vista científico. Si hubiera acceso a la *verdadera realidad*, la ciencia sería superflua y entonces ¿para qué molestarse con estos temas? ¿Por qué crear modelos si se puede acceder directamente al original? Obviamente, este no es el caso. Así pues, sigue siendo científicamente relevante si las teorías y los modelos pueden resistir una comprobación crítica de la realidad (por ejemplo, en forma de nuevos datos, poder explicativo o predictivo, así como potencial para derivar intervenciones, etc.) y se formulan de manera coherente. Para que la selección de criterios sea más comprensible, los resultados pueden analizarse con arreglo a distintos criterios, de modo que en lugar de un resultado selectivo surja una gama de resultados que, a su vez, refleje la variabilidad y complejidad del mundo. En el capítulo 6.3.1 tratamos en detalle la cuestión de la subjetividad en el contexto de la estadística de Bayes. Pero entonces hay que responsabilizarse de las decisiones, lo cual es más exigente que seguir ciegamente las convenciones.

¿Es ahora la subjetividad un defecto o un error en el proceso de investigación científica? Desde nuestro punto de vista, es una pregunta equivocada. Lo más importante es lo que se hace con la subjetividad. ¿Cómo se utiliza? ¿Qué nuevas posibilidades surgen de ello que no habrían sido posibles de otro modo? En la investigación cualitativa, a menudo es imposible reconstruir adecuadamente el objeto de estudio sin la experiencia subjetiva y la reflexión de los investigadores. No en vano, la APA (American Psychological Association) ha publicado un importante artículo que por fin (¡!) permite documentar adecuadamente las circunstancias especiales de la investigación cualitativa en relación con los artículos de revistas (Levitt, Bamber, Creswell, Frost, Josselson, & Suárez-Orozco, 2018). Con un retraso de unos 10 años, llegará a la zona de habla alemana.

Ya se ha completado una visión general de la teoría de Fisher y de la teoría Neyman-Pearson de la estadística clásica. El núcleo de la estadística clásica es (todavía) la prueba de significancia. La siguiente sección está dedicada a ésta y sus variantes, incluido el aspecto ritual. Un ritual es generalmente una acción regular que tiene un alto valor simbólico. Puede ser de naturaleza religiosa o secular, está culturalmente integrada y describe una comunicación regulada con el entorno (Schmidt, 1991). Y esa es precisamente la cuestión.

4.3.8 El proceso de una prueba estadística: El ritual del cero

> "El ruido de los ritos externos desaparece cuando surge la verdadera realización"
>
> (Ramakrishna, 1836-1886, místico indio).

La estadística clásica aún no ha llegado al punto descrito por Ramakrishna. Aquí sigue prevaleciendo el ritual de la prueba de hipótesis nula. El núcleo de la estadística clásica es la prueba estadística para determinar la significación estadística. Consiste en comparar un parámetro que representa una propiedad de los datos (= valor *p*) con una cantidad que en realidad debe especificarse teóricamente, el umbral crítico de significación. En la práctica, parece que nadie justifica teóricamente esta variable comparativa, sino que la adopta ciegamente qua habitus y expectativas generales. El proceso sigue un procedimiento fijo. Gigerenzer (2004b) habla aquí del ritual nulo o prueba de significación de hipótesis nula, o NHST para abreviar (ibid., p.588):

> „The null ritual:
>
> 1. Set up a statistical null hypothesis of 'no mean difference' or 'zero correlation.' Don't specify the predictions of your research hypothesis or of any alternative substantive hypotheses.
> 2. Use 5 % as a convention for rejecting the null. If significant, accept your research hypothesis. Report the result as $p < 0.05$, $p < 0.01$, or $p < 0.001$ (whichever comes next to the obtained p-value).
> 3. Always perform this procedure."

Esto se parece a lo siguiente para una simple *prueba t de dos muestras*. El código R para ello se encuentra en el archivo `ptII_quan_classicstats_nullritual.r`:

```
>   # define level of significance
>   # in accordance to lobbyism and bad habits >     crit.sig <- 0.05
>   # create some data
>   a <- c(1,1,4,7,3,6,2,5,3,3)
>   b <- c(6,9,4,8,4,3,12,1,11,6)
> ttest.ab.res <- t.test(a,b)
> ttest.ab.res
Welch Two Sample t-test
data: a and b
t = -2, df = 14, p-value = 0.04
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-5.673 -0.127
sample estimates:
mean of x  mean of y
3.5        6.4
> test.sig.NULLritual(p.value=ttest.ab.res$p.value, crit.sig=crit.sig)
WELCOME TO THE NULL RITUAL
The test against NULL (= NHST) is statistically significant
with p < 0.05.
You can reject H0.
```

y para el resultado estadísticamente no significativo de una prueba de correlación

```
cor.ab.res <- cor.test(a,b) R-Code
cor.ab.res
test.sig.NULLritual(p.value=cor.ab.res$p.value, crit.sig=crit.sig)
```

Gigerenzer (ibid., p.590) describe el proceso según el tardió Fisher (~1956/1973) en relación con las pruebas de significación de hipótesis nulas de la siguiente manera:

> *„Fisher's null hypothesis testing:*
> 1. Set up a statistical null hypothesis. The null need not be a nil hypothesis (i.e., zero difference).
> 2. Report the exact level of significance (e.g., p = 0.051 or p = 0.049). Do not use a conventional 5 % level, and do not talk about accepting or rejecting hypotheses.
> 3. Use this procedure only if you know very little about the problem at hand."

La repetición de la prueba t anterior es diferente según el tardió Fisher

```
THE LATE FISHER
The test resulted in an exact p-value of p = 0.0416009661836099.
Now hopefully you have learned something about
the research study, the design, the data, and your theory ---
Use your brain!
```

Y para la teoría de la decisión de Neyman-Pearson el procedimiento es el siguiente (Gigerenzer, ibid., p.590f.):

> *„Neyman–Pearson decision theory:*
> 1. Set up two statistical hypotheses, H_1 and H_2, and decide about α, β, and *sample size* before the experiment, based on subjective cost-benefit considerations. These define a rejection region for each hypothesis.
> 2. If the data falls into the rejection region of H_1, accept H_2; otherwise accept H_1. Note that accepting a hypothesis does not mean that you believe in it, but only that you act as if it were true.
> 3. The usefulness of the procedure is limited among others to situations where you have a disjunction of hypotheses (e.g., either $\mu_1 = 8$ or $\mu_2 = 10$ is true) and where you can make meaningful cost-benefit trade-offs for choosing alpha and beta."

Neyman-Pearson es bastante más amplio en su aplicación. En primer lugar, el tamaño de la muestra se calcula a priori con un análisis de potencia

```
> # Neyman-Pearson
> digits <- 2
> alpha <- 0.05
> seed <- 9876
> set.seed(seed)
> aprioripw.N <- power.t.test(n=NULL, delta=1.6, sd=2,
+          sig.level=alpha, power=0.8,
+          type=c("two.sample"),
+          alternative=c("two.sided"))
> aprioripw.N
Two-sample t test power calculation
n = 25.5
delta = 1.6
sd = 2
sig.level = 0.05
power = 0.8
alternative = two.sided
NOTE: n is number in *each* group
```

En consecuencia, generamos dos muestras. A continuación, podemos examinar los datos de la muestra empírica, por ejemplo, mediante la prueba *t* y la potencia de efecto *d* de Cohen.

```
a <- round(rnorm(n=25, mean=6.5, sd=2))
b <- round(rnorm(n=25, mean=7.8, sd=2))
describes(data.frame(a,b))
t.test(a,b)
DiM <- mean(a)-mean(b)
DiM
cohensd(b,a)
```

En general, el procedimiento según Neyman-Pearson resulta entonces de tal manera que, en contraste con el último Fisher el valor *p exacto* carece de interés, sino sólo la cuestión de si el valor *p* es menor o mayor que el *nivel de tasa de error* seleccionado.

```
> test.sig.NP(p.value=ttest.ab.res$p.value, alpha=alpha)
NEYMAN-PEARSON DECISION THEORY
The test decides based on a priori power analysis:
Two-sample t test power calculation
    n = 25.5
    delta = 1.6
    sd = 2
    sig.level = 0.05
    power = 0.8
    alternative = two.sided
NOTE: n is number in *each* group
Hypothesis testing according to Neyman-Pearson
Cohen's delta = -0.62
alpha = 0.05
p < alpha = TRUE
NOTE:
p < alpha, therefore accept H2.
```

La decisión de la prueba a favor de H_1 o H_2 se ejecuta en consecuencia. Con estos tres procedimientos se ha cubierto una gran parte de todos los artículos científicos (sociales) de las últimas décadas, siempre que no se hayan realizado de forma bayesiana o puramente exploratoria. También debe quedar claro qué se considera generalmente científico y qué se publica – a saber, lo que es estadísticamente significativo. El NHST descrito anteriormente es el principal, ya que muy pocos artículos tienen un procedimiento estrictamente según Fisher o Neyman-Pearson.

Así pues, hay tres procedimientos distintos que difieren sustancialmente entre sí. Para todas ellas es cierto que pertenecen a la prueba de significancia y que prácticamente ningún artículo de revista indica exactamente a qué versión pertenecen los valores notificados. Así, el ritual Cero del NHST, que todavía se practica ampliamente en la actualidad, difiere tanto del enfoque de Fisher como del de Neyman-Pearson en aspectos

fundamentales. *Ni Fisher ni Neyman-Pearson habrían de acuerdo con la NHST*. Los pioneros de la estadística clásica, por lo demás hostiles, probablemente habrían estado de acuerdo (Jaynes, 2003). Aparte de eso, ahora hay una serie de artículos muy críticos con el NHST (por ejemplo, Cohen, 1994; Gigerenzer, 2004b; Wagenmakers, 2007a; McShane, Gal, Gelman, Robert, & Tackett, 2019; y muchos más) y contraargumentos mucho menos elaborados (por ejemplo, Robinson & Wainer, 2001), que no necesariamente favorecen el propio NHST, sino que se refieren a alternativas clásicas como los intervalos de confianza o la prueba de Neyman-Pearson (Mayo, 1981, 2018) o simplemente desean reducir el valor p como criterio de significación estadística para que se adapte mejor, entre otras cosas, a los factores de Bayes (Benjamin et al., 2018-01-01). Greenwald, González, Harris y Guthrie (1996), por otra parte, exploran la cuestión de por qué las pruebas de hipótesis nulas siguen estando tan extendidas. Oevens (2012) utiliza el ejemplo del artículo NHST-crítico de Cohen (1994) para abordar aspectos de comprender correctamente las estadísticas y comunicarlas con precisión. Al hacerlo, intenta corregir opiniones defectuosas sobre NHST. Senn (2001), por su parte, aboga por los p-valores, pero advierte contra su uso poco realista y la idea de negar métodos alternativos de análisis.

El procedimiento básico se aplica con relativa independencia del nivel de escala, el tipo de método de análisis de datos, etc. – y por cierto, nadie trabaja realmente según Fisher o Neyman-Pearson. Pearson (para un resumen de las diferencias entre estos dos enfoques, cf. capítulo 4.5). Aunque los métodos de trabajo de Fisher y Neyman-Pearson difieren mucho, coincidieron en que el análisis estadístico no debe automatizarse. Ambos habrían rechazado vehementemente el ritual nulo automatizado. Tanto Fisher como Neyman-Pearson pensaban en algo al formular sus planteamientos. Como el anterior muestran, ambas teorías estadísticas tienen su legitimidad en diferentes contextos de aplicación.

La *ventaja* de este ritual Cero fijo es que los principiantes ostensiblemente pueden obtener resultados si conocen el nivel de escala de los datos, el problema (por ejemplo, diferencias entre dos grupos) y, posteriormente, el método de análisis estadístico en cuestión y pueden aplicarlo en el programa informático.

Pero, ¿tiene esto sentido, porque en la práctica significa que hay que ser capaz de pulsar botones con el ratón en una interfaz gráfica y confiar en los resultados, es decir, en la significación. Este tipo de estadísticas de libro de cocina no aprovecha todas las posibilidades y conduce a una situación en la que, en el peor de los casos, no se trata los datos conforme a las necesidades del caso concreto. Esto no sólo es problemático desde el punto de vista de la estadística bayesiana, sino también desde la desde el punto de vista de la estadística clásica moderna, que no funciona necesariamente según Bayes (por ejemplo, Gelman & Hill, 2007; Eid, Gollwitzer & Schmitt, 2010). Como ya mencionado, la estadística bayesiana también puede utilizarse mal en el sentido del ritual nulo (Gigerenzer & Marewski, 2015),si qua convención se erigen barreras que los factores de Bayes deben alcanzar para que un resultado sea tomado en serio y publicado. En tal caso, la diferencia entre Bayes y la estadística clásica se reduce a cero. Expulsar al diablo con Belcebú es un intento desfavorable de solución y demuestra pocos conocimientos adquiridos sobre la base de las últimas décadas.

La *desventaja* del NHST es que el ritual Cero no permite un enfoque adaptado a cada caso. En la práctica, suelen predominar las rutinas y los supuestos inconscientes, como explican detalladamente Gelman y Loken (2013) utilizando el ejemplo del muy discutido estudio de Bem (2011a) sobre la clarividencia. Los investigadores también tienen que lidiar con el hecho de que la naturaleza humana busca la confirmación de las propias suposiciones en lugar de cuestionarlas y ponerlas a prueba de forma crítica. Esto es, sin embargo, lo que exige Popper, quien desgraciadamente aplica esta lógica, por ejemplo, a la conocimiento inductivo y la estadística bayesiana, pero no se lo aplicó a sí mismo. El problema no es originalmente cuantitativa. Así pues, el análisis de secuencias en el marco de la Hermenéutica Objetivo (véanse los capítulos 11.2 y 11.9) exige con vehemencia un examen crítico de sus propios presupuestos y enmarca todo el proceso precisamente en torno a este punto crítico. Teniendo en cuenta la tendencia a confirmar más que a cuestionar, la falta de estudios de replicación no es realmente sorprendente. Aparte de eso, hay una falta de aprecio general por ella en la ciencia, que tiende a ser operacionalizado en términos de dólares de investigación y número de publicaciones, no sobre la exhaustividad y la estabilidad a largo plazo de los resultados.

Programas informáticos como SPSS® de IBM, que prefieren los clics del ratón a la planificación basada en scripts – aunque sigue siendo fácilmente posible – dificultan el tratamiento correcto de las estadísticas. Por supuesto, incluso los principiantes obtienen resultados de esta manera, pero sigue sin estar claro si después se ha realizado e interpretado correctamente un análisis. Se vuelve peligroso cuando un software

como produce resultados estadísticos inferenciales que, estrictamente hablando, tienen que ser calificados de erróneos. Esto incluye la salida "Significación = 0,000***". En primer lugar la importancia no la determina un programa informático, sino un criterio teórico establecido de antemano por un ser humano, ya que existe una diferencia entre los valores p y las barreras de significación. Y las probabilidades con exactamente p = 0 no existen – y si existieran, no se podría confiar en ellas. El hecho de que probablemente haya un problema de redondeo en la presentación es secundario. Muy pocas personas utilizarán la línea de comandos de SPSS® e introducirán el código directamente o comprueban los decimales. Por desgracia, así lo demuestran años de experiencia en supervisión y asesoramiento de trabajos de cualificación científica. Aquí es donde se nota los puntos fuertes de programas informáticos como R (2019d). Trabajar con R es más difícil al principio, pero después es mucho más seguro que los resultados sean los que el usuario planificó e introdujo en el análisis. Esto no protege contra el uso indebido, incluso en R.

Tras describir el procedimiento básico de la prueba de significancia, la siguiente sección examina en detalle cómo se toma realmente una decisión sobre una prueba. Esto conduce a la probabilidad de los datos bajo la validez de la hipótesis nula, conocido como valor p.

4.3.9 La probabilidad de los datos como base de las decisiones sobre las pruebas

> „If *P* is small, that means that there have been unexpectedly large departues from prediction. But why should these be stated in terms of *P? [...] What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred.* This seems a remarkable procedure. On the face of it, the evidence might more reasonably be taken as evidence for the hypothesis, not against it. The same applies to all current significance tests based on *P* integrals“ (Sir Harold Jeffreys, 1939/1961, p.385, cursiva en original, nota de los autores)

Entonces, ¿qué examina realmente una prueba estadística clásica y en qué se basa su razonamiento para justificar las decisiones? La cita de Jeffreys lo formula con un fino inglés humor que no necesita más interpretación, pero que debe ser tomada en serio. Tschirk (2014, p.80) comenta con bastante sobriedad la prueba estadística clásica:

> "Una prueba estadística no nos dice si una hipótesis es cierta. Si la variable de prueba entra dentro del intervalo de aceptación, puede tratarse de una coincidencia: por casualidad, la muestra se ajusta bien a la hipótesis, aunque ésta sea falsa. Si la variable de prueba cae dentro del rango de rechazo, también puede tratarse de una coincidencia: por casualidad, la muestra se ajusta mal a la hipótesis, aunque ésta sea correcta. Una prueba estadística clásica tampoco indica la probabilidad de que una hipótesis sea cierta. Desde un punto de vista objetivista, no existe probabilidad alguna, porque la hipótesis es cierta, sólo que no lo sabes. La pregunta sólo tiene sentido si se piensa de forma subjetiva y además se atribuye una probabilidad a los acontecimientos cuya ocurrencia o no ocurrencia es cierta. Pero incluso así, la prueba no proporciona la probabilidad que buscamos. Lo justificamos con el teorema de Bayes".

La razón de la incapacidad de las pruebas clásicas para hacer afirmaciones sobre la probabilidad de las hipótesis es, pues, la siguiente (véase también el teorema de Bayes, cap. 6.4): La base del ritual Cero y, por tanto, de todas las decisiones sobre hipótesis en estadística clásica reside en el cálculo de la probabilidad condicional de los datos disponibles o más extrema dada la hipótesis Nula formulada más o menos (in)específica (= valor p).

La Figura 4.23 muestra una distribución Nula simulada de las medias y los dos rangos de rechazo en el caso de una prueba de dos caras, es decir, para la prueba de una muestra y la prueba de dos muestras. Se parte del supuesto de una distribución normal. El código R no se imprime (`ptII_quan_classicstats_NHST_nulldist.r`).

Hay que añadir la frase "o más extremo" porque el área de rechazo es un área bajo una curva (véase la Fig. 4.23) y no un punto singular. El área bajo una curva puede interpretarse de forma probabilística. El área

de rechazo α designa un límite superior (= máximo) de la probabilidad de error y, por tanto, *incluye todos los valores por debajo* de los cuales la probabilidad es inferior α. El rango de rechazo del H_0 es, por tanto, una superficie y no un punto. Por otra parte, se supera el umbral crítico de significación, que distingue entre significativo y no significativo. Se trata de un límite con una extensión de prácticamente cero, como $p < 0.056$, $p < 0.00123$, etc., por utilizar niveles no convencionales como ejemplo. El valor p da a su vez la probabilidad de todos aquellos datos (observados, no observados) que tienen el valor p como límite superior. Se trata, en principio, de un conjunto infinitamente grande de posibles conjuntos de datos e incluye tanto conjuntos de datos observados como no observados, es decir, desconocidos.

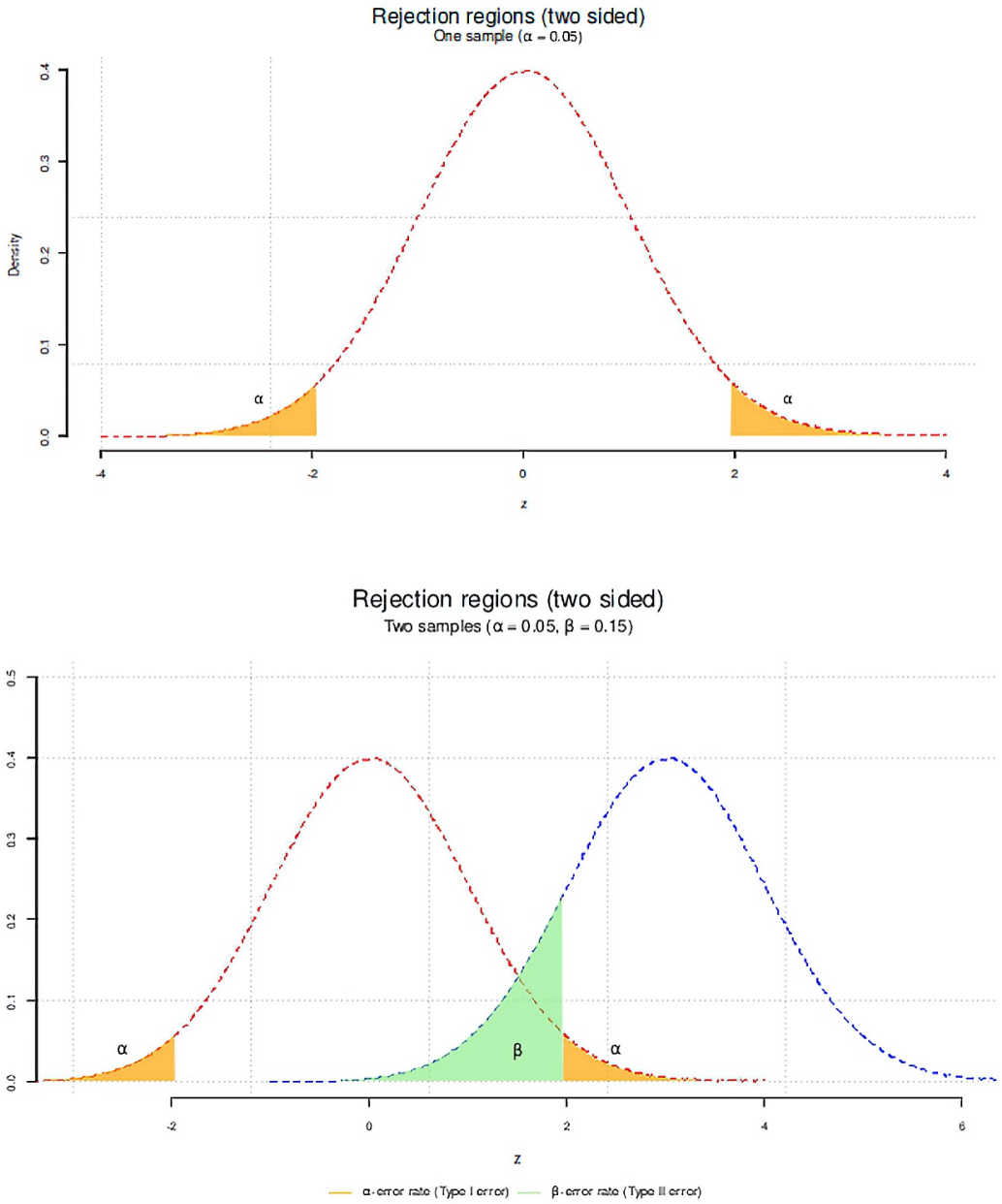


Figura 4.23. *Valor p (una y dos muestras, prueba bilateral)*

Por el contrario, la barrera de significancia (véase la Fig. 4.23) es una línea vertical con extensión cero. Esto simplemente marca la transición de significativo a no significativo, pero no contiene ningún área bajo la curva y, por lo tanto, no representa un conjunto de datos (empíricos, simulados, imaginarios). Esta diferenciación es importante. El umbral de significación es una magnitud teórica, no empírica. Como tal,

debe tener una base teórica, no empírica. En la práctica no es así. En la investigación, esta barrera no suele justificarse en absoluto y se aplica qua convención y sin referencia al contenido.

Una vez más, la probabilidad de error establecida para rechazar falsamente la hipótesis nula es un límite superior predefinido hasta el cual el error no se desea, pero se tiene en cuenta.

4.3.9.1 *Cálculo del valor p*

El valor p se define como

$$\text{Valor p} = \text{p(datos empíricos o más extremos} \mid H_0) \qquad (4.16).$$

Es el resultado de un algoritmo de prueba, es decir, el procedimiento estadístico de prueba (por ejemplo, la prueba *t*, la prueba χ^2, los predictores en un modelo lineal, etc.), que a su vez genera una variable de prueba, que a su vez tiene una ubicación exacta en la distribución de la hipótesis Nula – si ésta es válida o verdadera. Esta localización da lugar directamente al valor *p* como probabilidad sobre la base del conocimiento de la distribución exacta de la hipótesis Nula. Para determinar el valor *p exacto*, nos fijamos en dónde se sitúa la variable de prueba empírica en la distribución de la hipótesis Nula y qué área abarca. Esta zona corresponde a una probabilidad y corresponde al valor *p*. La distribución de la hipótesis Nula se genera a partir del conocimiento de el procedimiento de la prueba estadística y varía según los procedimientos analíticos. Así pues, parece que R.A. Fisher estaba intuitivamente muy dotado para encontrar las distribuciones de prueba correctas para diferentes procedimientos de prueba estadística (por ejemplo, el análisis de la varianza), como señalan Jaynes (2003) y Salsburg (2001).

Dado que el área bajo una curva siempre puede interpretarse en términos probabilísticos, nos fijamos en la probabilidad *p* con la que se localiza la variable de prueba y, en función de si la prueba es unilateral o bilateral, qué parte o partes del área están cubiertas. Dependiendo de la distribución (de mezcla), las probabilidades pueden variar drásticamente.

Para ilustrar este problema, veamos un ejemplo en el que números aleatorios de un de una distribución normal estándar, por un lado, y números aleatorios de una normal mixta, por otro, se representan en forma de gráfico de densidad (véase la Fig. 4.24). La distribución normal mixta consiste en una distribución normal estándar a la que el 20 % de los valores de a distribución normal con media $\mu = 0$ y desviación estándar $\sigma = 5$ se añaden (`ptII_quan_classicstats_pvalue-as-base.r`).

```
# normal vs. mixed normal distribution
seed <- 556
set.seed(seed)
n <- 1e5
fac <- .2
SD <- 5
p1 <- rnorm(n)
p1.5 <- rnorm(n*fac, sd=5)
p2 <- sample(c(p1,p1.5), size=n, replace=FALSE)
par(oma=c(2,1,3,1), "cex.axis"=1, bty="l")
hist(p1, panel.first=grid(), prob=TRUE, border="white", col="skyblue",
           main="", xlab="quantile", ylab="Density")
lines(density(p1), col="darkred", lwd=2, lty=2)
lines(density(p2),col="black", lwd=2, lty=3)
# legend
legend("topright", legend=c("NV","NV + t"),
lwd=c(2,2), lty=c(2,3), col=c("darkred","black"), bty="n")
mtext("Histogram", 3, line=4, cex=2)
mtext("normal + mixed normal distribution", 3, line=2, cex=1.5)
# quantiles
probs <- c(0,0.01,0.025,0.05,0.1,0.25,0.5,0.75,0.9,0.95,0.975,0.99,1)
q1 <- quantile(p1, prob=probs)
q2 <- quantile(p2, prob=probs)
```

Como puede observarse, las colas de las distribuciones difieren significativamente entre sí, lo que en la práctica da lugar a valores *p* diferentes y puede comprenderse fácilmente de forma numérica a partir de los cuantiles y las estadísticas descriptivas. La distribución normal mixta es más robusta y conservadora en las colas que la distribución normal pura, es decir, muestra valores de densidad mayores para el mismo valor de cuantil. Por el contrario, la distribución normal pura muestra densidades más altas en el rango medio que la distribución mixta. Sin embargo, hay que "ver" que la distribución es mixta. distribución. En la práctica, esto es difícil o incluso imposible sin información previa. Sin embargo, como puede verse, tiene consecuencias cuando se trata de pruebas estadísticas basadas en valores *p*.

```
> cbind(q1,q2)
         q1        q2
0%     -4.29942 -20.65765
1%     -2.32074  -7.90240
2.5%   -1.95310  -5.18947
5%     -1.63761  -2.81184
10%    -1.27826  -1.67580
25%    -0.66931  -0.79014
50%     0.00273   0.00493
75%     0.67378   0.80180
90%     1.28098   1.68805
95%     1.64028   2.87094
97.5%   1.95804   5.26883
99%     2.33842   7.85743
100%    5.29746  19.31861
> q2/q1
0%   1%   2.5% 5%   10%  25%  50%  75%  90%  95%  97.5% 99%  100%
4.80 3.41 2.66 1.72 1.31 1.18 1.81 1.19 1.32 1.75 2.69  3.36 3.65
> apply(cbind(q1,q2),2,function(x) c(summary(x),sd=sd(x),var=var(x)))
         q1        q2
Min.    -4.29942 -20.65765
1st Qu. -1.63761  -2.81184
Median   0.00273   0.00493
Mean     0.07948  -0.09359
3rd Qu.  1.64028   2.87094
Max.     5.29746  19.31861
sd       2.49195   9.13686
var      6.20980  83.48213
```

La relación entre las dos distribuciones se muestra en el diagrama de dispersión (véase la Fig. 4.25).

```
# scatterplot R-Code
plot(sort(p1),sort(p2), type="l", col="darkred", bty="n",
           pre.plot=grid(), xlab="standard normal distribution",
           ylab="mixed normal distribution")
abline(lm(sort(p2) ~ sort(p1)))
mtext("Relationship standard normal dist. and mixed normal dist.",
           side=3, line=2, cex=2)
```

Formulada como un proceso, la prueba de significación funciona del siguiente modo:

1. Establecer el exceso de probabilidad crítica que determina si una cantidad estadísticamente interesante pasa a ser significativa o no. Esta elección debe hacerse teniendo en cuenta consideraciones teóricas. En la práctica, por desgracia, suelen ser las convenciones y no la adecuación al contexto las que deciden, por ejemplo, 0,05, 0,01, 0,001, etc. Además, hay que determinar si se trata de una prueba direccional unilateral o no, direccional bilateral, la elección del procedimiento de análisis de datos, etc.
2. Establecer la distribución de la H_0. Esto puede hacerse conociendo exactamente el aspecto de la distribución Nula de la variable de prueba estadística de interés (por ejemplo, el valor *t*). Por ejemplo, la prueba *t* produce una variable de prueba distribuida *t* con los grados de libertad correspondientes, en el caso más sencillo "tamaño de la muestra menos uno", si no se estiman otros

parámetros en el modelo. Otra posibilidad es simular la distribución de la prueba utilizando los datos empíricos. Este es el enfoque adoptado en el ejemplo anterior de la Figura 4.23.

3. Calcular la variable empírica de prueba cuya distribución corresponde a la hipótesis según la distribución de H_0. Si la distribución Nula es aproximadamente la distribución de Poisson, pero la variable de prueba de interés es la distribución gamma. El cálculo de la variable de prueba empírica se realiza según el procedimiento de prueba estadística elegido (por ejemplo, prueba *t*, valores *t* en regresión, valor χ^2 en prueba χ^2, etc.). La variable de prueba empírica procede directamente de los datos recopilados y es el resultado de aplicar a los datos el procedimiento de análisis de datos (por ejemplo, prueba *t*, regresión, prueba χ^2, etc.). datos.
4. La variable de prueba resultante se sitúa en la distribución Nula. Esto significa que se examina en qué cuantil de la distribución nula se localiza la variable calculada empíricamente. En caso de se realiza mediante código R. En el caso de una simulación, esto se hace mediante la función `cuantil`. El resultado es el valor *p exacto* como propiedad directa de los datos.
5. En consecuencia, se toma una decisión según el tipo de prueba de significación (véase el capítulo 4.3.8), dependiendo de si el valor *p* es menor, igual o mayor que el exceso de probabilidad crítica. Prácticamente nunca los valores son exactamente igual.

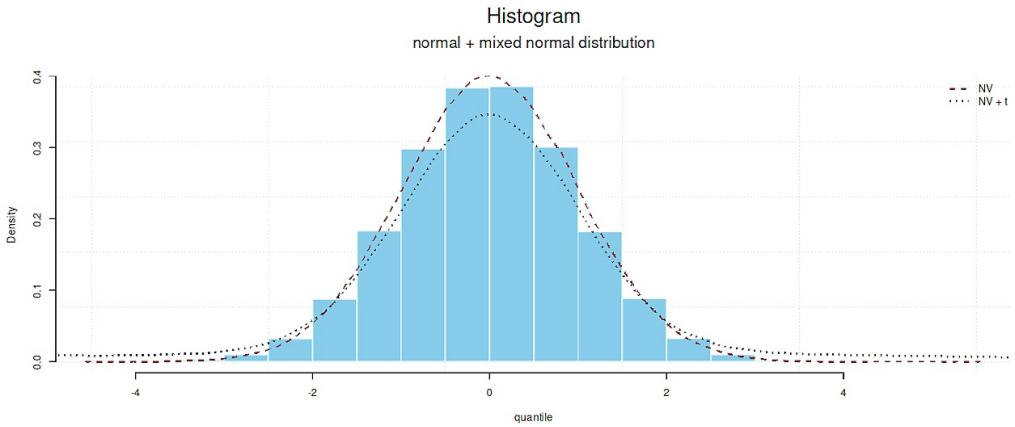


Figura 4.24. *Diagrama de densidad (distribución normal, distribución normal mixta)*

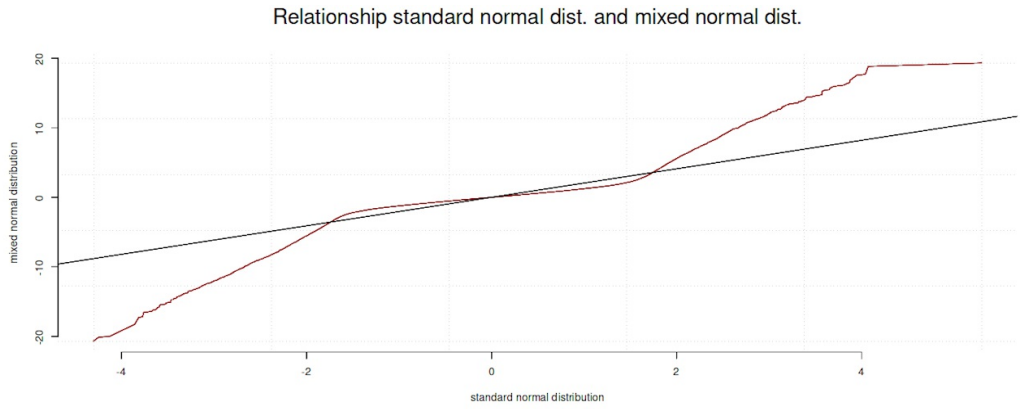


Figura 4.25. *Diagrama de densidad (distribución normal, distribución normal mixta)*

La figura 4.23 muestra este proceso a lo largo de la simulación de una distribución nula. En primer lugar, para $\mu = 100$ y $\sigma = 10$, se generan mediante simulación $n = 10\,000$ valores medios empíricos, cada uno con un tamaño de muestra $n = 30$, y se calcula la densidad de esta distribución media (`ptII_quan_-classicstats-pvalue-as-base.r`):

```
# calculation of the p value
# simulation H0
seed <- 9876
set.seed(seed)
mu1 <- 100
sigma1 <- 10
trials <- 10000
n <- 30
mean.sim <- replicate(trials, mean(rnorm(n=n, mu1, sigma1)))
mean.sim.dens <- density(mean.sim)
```

El conjunto se representa gráficamente incluyendo las áreas de rechazo (área transparente en la Fig. 4.23) para un nivel convencional $\alpha = 0.05$ con probabilidades críticas de superación para una prueba de dos caras. La función de cuantiles `quantile()` se utiliza para determinar las probabilidades críticas de superación:

```
# critical values R-Code
alpha <- 0.05
crit.sig.values <- quantile(mean.sim, probs=c(0.025, 0.975))
# histogram and density
color <- rgb(1,0,0,alpha=.2)
xlim <- range(mean.sim.dens$x)
ylim <- c(0,max(mean.sim.dens$y))
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l")
hist(mean.sim, ylim=ylim, panel.first=grid(), prob=TRUE, border="white",
           col="skyblue", main="", xlab="mean values", ylab="Density")
lines(mean.sim.dens, col="red", lwd=2)
# lower limit rejection area
polygon(x=c(xlim[1],
mean.sim.dens$x[mean.sim.dens$x < crit.sig.values[1]],
           crit.sig.values[1]), y=c(ylim[1],
           mean.sim.dens$y[mean.sim.dens$x <= crit.sig.values[1]],
           ylim[1]), col=color, border=NA)
# upper limit rejection area
polygon(x=c(crit.sig.values[2],
           mean.sim.dens$x[mean.sim.dens$x > crit.sig.values[2]],xlim[2]),
           y=c(ylim[1],
           mean.sim.dens$y[mean.sim.dens$x >= crit.sig.values[2]],
           ylim[1]), col=color, border=NA)
mtext("Simulation H0 (null distribution)", outer=TRUE,
           line=-2, cex=1.5, side=3)
```

Seleccionamos ficticiamente un valor empírico de $x = 104$ como valor medio de los datos empíricos. A continuación se calcula el valor p:

```
# now arbitrary empirical value
xbar1 <- 104
# calculate p-value
mean.sim.sort <- sort(mean.sim)
```

para ambos lados de la zona de rechazo:

```
> # upper limit
> emp.pv <- 1-length(mean.sim.sort[which(mean.sim.sort < xbar1)])/trials
> emp.pv
[1] 0.0151
> # lower limit
> 1 - length(mean.sim.sort[which(mean.sim.sort > xbar1)])/trials
[1] 0.985
```

Ahora se calcula el intervalo crítico en función del α-nivel seleccionado:

```
> # calculate critical values for significance test
> mean.sim.sd <- sd(mean.sim)
> fak <- qnorm(1-alpha/2)
> crit <- mu1 + c(-fak,fak)*mean.sim.sd
> crit
[1] 96.4 103.6
```

La figura debe también mostrar el valor empírico y una leyenda:

```
# abline(v=crit, col="magenta", lty=2, lwd=1)
lines(x=c(xbar1,xbar1), y=c(0,max(mean.sim.dens$y)/2), col="darkred", lty=3, lwd=3)
# legend
legend("topright",
           legend=c(paste("crit. (low) = ",round(crit[1],2),sep=""),
           paste("crit. (up) = ",round(crit[2],2),sep=""),
           eval(substitute(expression(paste(bar(x)," (emp.) = ",
           xbar1,sep="")),list(xbar1=xbar1)))),
           lwd=c(2,2,2), lty=c(1,1,3), col=c(color, color,"darkred"),
           bty="n")
mtext(eval(substitute(expression(paste(mu," = ",mu1,
           " | ",sigma," = ",sigma1,
           " | ",alpha," = ",alpha1," | ",bar(x)," = ",xbar1,
           " | N = ",n," | trials = ",trials)),
           list(mu1=mu1, sigma1=sigma1, alpha1=alpha, xbar1=xbar1,
           n=n, trials=trials))), 1, line=5, cex=1.1)
```

La variante abreviada de la prueba de significación bilateral puede aplicarse fácilmente:

```
> # test xbar
> xbar1 > mean.sim.sort[trials*(1-alpha/2)] |
+          xbar1 < mean.sim.sort[trials*(alpha/2)]
[1] TRUE
```

Obviamente, la primera expresión es lógicamente VERDADERA y la segunda lógicamente FALSA, resultando un VERDADERO lógico a lo largo de la operación lógica AND. Dado que se trata de una prueba bilateral, para mantener la hipótesis Nula debería haber dado como resultado un FALSO lógico. Así pues, se rechaza la hipótesis Nula por no ser estadísticamente significativa al nivel α. Esto no significa que sea (prácticamente) relevante. También sería posible sin más comparar el valor *p* con el nivel medio α (porque es una prueba bilateral):

```
> # equivalent
> emp.pv < alpha/2
[1] TRUE
```

Esto conduce al mismo resultado. En el caso de que la distribución Cero se conozca con exactitud, esto permite calcular directamente las probabilidades críticas de superación. Si, por ejemplo, una variable tiene distribución *t*, la muestra es $n = 75$ y el nivel es $\alpha = 0.0412$, la prueba bilateral arroja las siguientes probabilidades de superación hacia la significación:

```
> # direct calculation
> # two-sided test
> # t distributed values
> n1 <- 75
> alpha <- 0.0412
> c(-1,1)*qt(1-alpha/2, df=n1-1)
[1] -2.08 2.08
```

Si la variable de prueba estuviera distribuida normalmente y el nivel α convencionalmente $\alpha = 0.05$, resultaría el valor conocido de la bibliografía:

```
> # normal distribution
```

```
> alpha <- 0.05
> c(-1,1)*qnorm(1-alpha/2)
[1] -1.96 1.96
```

que, redondeado, da 1.96. Debido a la distribución t ligeramente diferente, esto corresponde bajo condiciones t-distribuidas a

```
> c(-1,1)*qt(1-alfa/2, df=n1-1)
[1] -1.99 1.99
```

Esto da como resultado un valor de 1.99, es decir ≈ 2. Gelman y Hill (2007) dan este valor de $t \approx 2$ para modelos lineales, que puede considerarse una pauta aproximada (es decir, una "regla empírica") para dejar un parámetro en el modelo – siempre que los supuestos teóricos, la dirección y la magnitud del efecto asociado son correctos y no se aplica de forma dogmática (véase el capítulo 4.3.3.2). Se trata de una directriz, no de una norma exacta ni de una declaración tipo ley. Para examinar el comportamiento asintótico, basta con fijar los grados de libertad en infinito (= infinidad, realizada en R con `Inf`):

```
> c(-1,1)*qt(1-alfa/2, df=Inf)
[1] -1.96 1.96
```

Para comprobar si esto es realmente igual al caso asintótico, es decir, la distribución normal, se utiliza `all.equal()`:

```
> # test for roughly "the same"
> # (not necessarily 100% identical due to rounding)
> all.equal(c(-1,1)*qnorm(1-alfa/2), c(-1,1)*qt(1-alfa/2, df=Inf))
[1] TRUE
```

Los valores son los mismos. La variante presentada sólo funciona para distribuciones simétricas. Para los asimétricos, el intervalo de cuantiles superior e inferior debe calcularse por separado y no viene dada simplemente por c(-1,1)*qt(...) o c(-1,1)*qnorm(...). Un ejemplo es la distribución beta con los parámetros `shape1` y `shape2` 2 y 5:

```
# cola inferior
> a1 <- 2
> b1 <- 5
> qbeta(alpha/2, shape1=a1, shape2=b1, lower.tail=TRUE)
[1] 0.0433
# upper tail
> qbeta(alpha/2, shape1=a1, shape2=b1, lower.tail=FALSE)
[1] 0.641
```

Dado que la distribución con estos parámetros parece asimétrica, las diferentes probabilidades no son sorprendentes (véase la Fig. 4.26). Siempre hay que tener en cuenta la forma de la distribución en las pruebas clásicas, especialmente cuando se trata de calcular los intervalos de confianza (véase cap. 4.3.3.6).

El procedimiento descrito aquí refleja más o menos el que se utiliza en todas las pruebas estadísticas clásicas. Ahora, sin embargo, habría que preguntarse:

> Recordatorio 4.2: Probabilidades de interés
>
> ¿Por qué calcular una probabilidad condicional de los datos cuando lo que realmente interesa son las probabilidades de hipótesis contrapuestas o el posible valor añadido de integrar las hipótesis contrapuestas en un modelo más complejo?

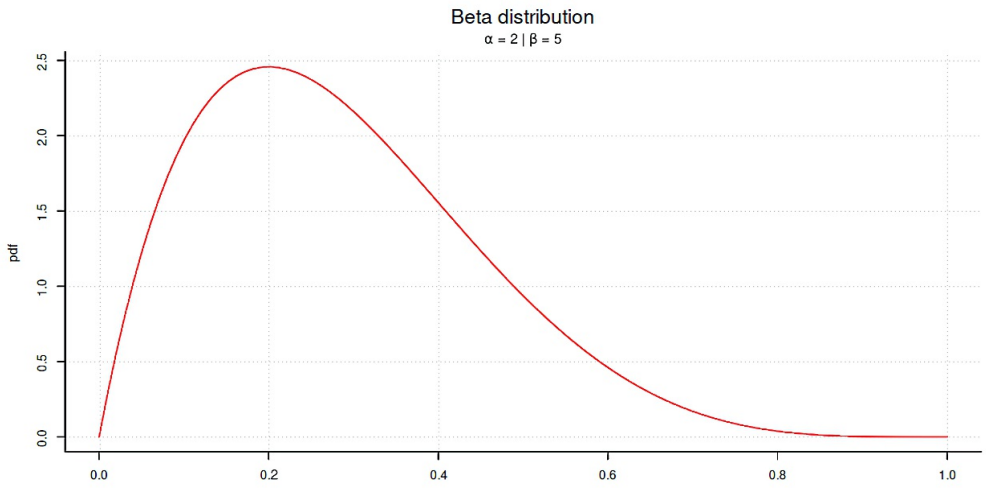


Figura 4.26. *Beta Verteilung (a = 2, b = 5)*

En la estadística clásica, no existe una inferencia causal lógica directa de la probabilidad de los datos (información) a la probabilidad de las hipótesis, porque eso sería la probabilidad posterior bayesiana. A título comparativo

$$\text{Valor } p = p(\text{Datos empíricos o más extremos} \mid \text{Hypotesis (Nula)}) \quad (4.17)$$

$$\text{posterior } p(\text{bayesiano}) = p(\text{Hypotesis} \mid \text{Información o datos empíricos}) \quad (4.18)$$

Uno no puede calcularse directamente a partir del otro: *La estadística clásica no puede calcular p(Hypotesis | Datos empíricos) en absoluto*, como puede verse en la cita anterior de Tschirk (2014) (véase al principio del capítulo 4.3.9). Este cálculo sólo es posible utilizando la estadística de Bayes. Sin embargo, esto introduce supuestos a los que la estadística clásica no es capaz de hacer frente. Entre ellas se incluyen las probabilidades a priori (reproche: subjetividad, véase también el capítulo 6.3.1) o la suposición de que los parámetros estimados son variables aleatorias con una distribución aleatoria cada una y no variables fijas (mensurables) propensas a errores. Además, el concepto de probabilidad no se define por la frecuencia relativa de los sucesos, sino que simplemente se puede asignar a los sucesos una probabilidad lo más fundada posible (O'Hagan, Buck, Daneshkhah, Eiser, Garthwaite, Jenkinson, Oakley & Rakow, 2006). Nadie pretende que este proceso sea trivial. Se cuestionan las investigaciones previas, las revisiones bibliográficas, los metaanálisis o los análisis contextuales. Los requisitos contextuales de un enfoque de este tipo, así como las orientaciones sobre el uso de información cualitativa, se detallan en O'Hagan et al. (ibíd.) y en la sección 6.14.6.

4.3.9.2 *Sobre la relación entre la distribución normal y la distribución t*

La figura 4.27 muestra la *distribución normal* y *la distribución t* con df = 1 a 10 grados de libertad. Como puede observarse, existe una clara diferencia entre la distribución *t* y la distribución normal, especialmente en las colas. Esta diferencia conduce naturalmente a valores *p* muy diferentes, especialmente con pocos grados de libertad, simplemente por la elección de la hipótesis de distribución. Como hemos visto anteriormente, al aumentar los grados de libertad, la distribución *t* se funde con la distribución normal y las diferencias tienden asintóticamente a cero. Por esta razón, la estadística clásica prefiere las condiciones asintóticas de infinito o las asume *per fiat* sin justificación empírica, ya que es mucho más fácil calcularlas bajo supuestos de distribución normal (véase la Fig. 4.27, `ptII_quan_classicstats_normal-vs-t.r`):

```
# relationship of normal and t distribution
# normal distribution and t distribution rejection area under H0
sek1 <- seq(-6,6,length.out=150)
dfree <- 2
norm.dens <- dnorm(sek1)
xlim.n <- range(sek1)
ylim.n <- c(0,max(norm.dens))
t.dens <- dt(sek1, df=dfree)
xlim.t <- range(sek1)
ylim.t <- c(0,max(t.dens))
# plot
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l")
plot(sek1, norm.dens, panel.first=grid(), type="l", col="red",
           bty="l", lwd=1, main="", xlab="", ylab="density")
lines(sek1, t.dens, col="steelblue", lwd=1, lty=2)
legend("topright", legend=c("normal",
           paste("t (df = ",dfree,")",sep="")),
           lty=c(1,2), lwd=1, col=c("red","steelblue"),
           bty="n", bg="yellow")
mtext("Normal versus t-distribution", outer=TRUE, line=-2,
           cex=1.5, side=3)
alpha <- 0.05
# critical regions
crit.nv <- c(-1,1)*qnorm(1-alpha/2)
crit.tv <- c(-1,1)*qt(1-alpha/2, df=dfree)
# histogram and density
color.n <- rgb(1,0,0,alpha=.2)
color.t <- rgb(70/255,130/255,180/255,alpha=.2)
# normal distribution
# lower limit rejection area
polygon(x=c(xlim.n[1],sek1[sek1 <= crit.nv[1]], crit.nv[1]),
y=c(ylim.n[1],norm.dens[sek1 <= crit.nv[1]], ylim.n[1]), col=color.n, border=NA)
# upper limit rejection area
polygon(x=c(crit.nv[2],sek1[sek1 > crit.nv[2]],xlim.n[2]),
y=c(ylim.n[1],norm.dens[sek1 >= crit.nv[2]],ylim.n[1]), col=color.n, border=NA)
# t distribution
# lower limit rejection area
polygon(x=c(xlim.t[1],sek1[sek1 <= crit.tv[1]], crit.tv[1]),
y=c(ylim.t[1],t.dens[sek1 <= crit.tv[1]], ylim.t[1]), col=color.t, border=NA)
# upper limit rejection area
polygon(x=c(crit.tv[2],sek1[sek1 > crit.tv[2]],xlim.t[2]),
y=c(ylim.t[1],t.dens[sek1 >= crit.tv[2]],ylim.t[1]), col=color.t, border=NA)
```

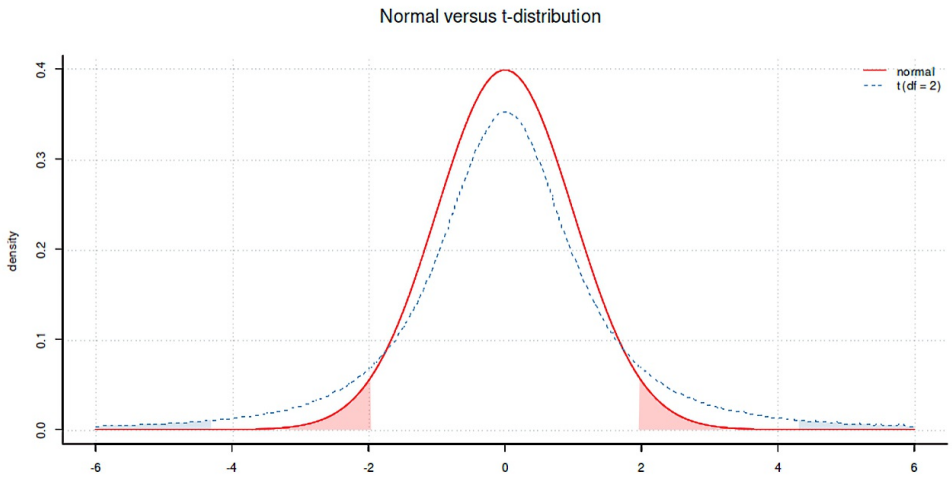


Figura 4.27. *Distribución normal vs. Distribución t*

En el presente caso, la diferencia entre los valores de la distribución *t* frente a la distribución normal en el eje X (= rango de rechazo) al mismo nivel α es de

```
>   # critical values n vs. t
>   # ratio seen from t perspective > crit.nv
[1] -1.96 1.96
> crit.tv
[1] -4.3 4.3
> crit.tv/crit.nv
[1] 2.2 2.2
```

Expresado en probabilidades, podemos ver

```
> # probs n vs. t
> pnorm(crit.nv)
[1] 0.025 0.975
> pt(crit.nv,df=dfree)
[1] 0.0945 0.9055
```

el conocido nivel α de dos lados para la distribución normal, pero valores fuertemente desviados para la distribución *t*. La diferencia

```
> pt(crit.nv,df=dfree)-pnorm(crit.nv)
[1] 0.0695 -0.0695
```

corresponde a una desviación de 0,0695 o ≈ 7 %, es decir, una diferencia sustancial según el contexto. Por lo tanto, una recomendación común es que para muestras pequeñas (n < 30) y, por lo tanto, con pocos grados de libertad, la distribución *t* es preferible a la distribución normal. Es más conservador y robusto – es decir, requiere valores *p* más pequeños – para identificar valores *p* estadísticamente significativos en el mismo nivel α. Las colas más grandes de la distribución (véase la Fig. 4.27) indican una mayor robustez de la *t* – frente a la distribución normal. Al aumentar el tamaño de las muestras, esta diferencia desaparece. Por ejemplo, a partir de df = 140 grados de libertad, la diferencia entre la distribución *t* y la normal ya es inferior a 0,001 (< 0,1 %) en el nivel α = 0,05, como se muestra en la Fig. 4.28:

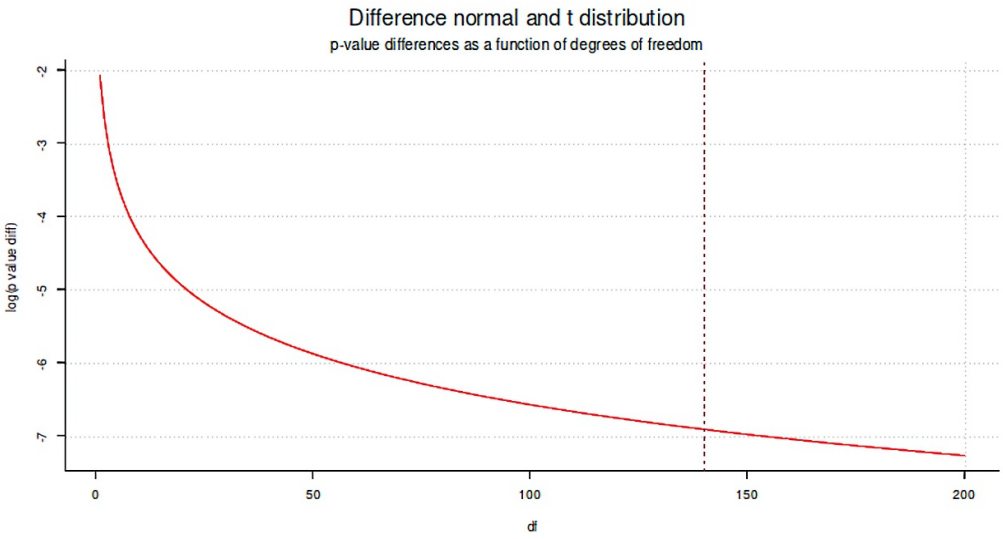


Figura 4.28. *Distribución normal frente a distribución t (dependiente de df's)*

```
cols <- terrain.colors(dfs)
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l")
plot(sek2, dnorm(sek2), panel.first=grid(), type="l", col="red",
```

```
          bty="l", lwd=2, main="", xlab="", ylab="Density")
for(i in 1:dfs) lines(sek2, dt(sek2, df=i), col=cols[i], lty=2)
legend("topright", legend=c("N",paste("t (",1:dfs,")",sep="")),
          lty=c(1, rep(2,dfs)), lwd=1, col=c("red",cols), bty="n", bg="yellow")
mtext("Normal versus t-distribution", outer=TRUE, line=-2, cex=1.7, side=3)
mtext("for various degrees of freedom", outer=TRUE, line=-3.6, cex=1.2, side=3)
```

4.3.9.3 *Excursus - Teorema del límite central*

La *distribución normal* desempeña un papel especial en estadística, que se basa en el *teorema del límite central*. Pólya (1920) fue el primero en formularla. En general, los teoremas del límite central establecen que las sumas de variables aleatorias independientes a medida que el tamaño de la muestra aumenta asintóticamente convergen asintóticamente hacia una distribución estable. Si las variables aleatorias tienen una varianza finita y positiva, su suma converge hacia la distribución normal. Existen diferentes fórmulas para ello. Una variante muy común supone variables aleatorias independientes e idénticamente distribuidas con valor esperado finito y varianza finita según Jarl Waldemar Lindeberg (1876-1932) y Paul Lévy (1886-1971), de modo que las variables de la suma pueden considerarse aproximadamente normalmente distribuidos. En otra formulación, la distribución idéntica no se considera un requisito previo necesario y se imponen otras condiciones. para que ninguna de las variables pueda ejercer una influencia demasiado grande en la distribución (condición de Lindeberg; Lindeberg, 1922; condición de Lyapunov, Weisstein, s.f.).

Con tamaños de muestra finitos, debe examinarse en cada caso si la convergencia a la distribución normal procede de manera uniforme. Según el teorema de Berry-Esseen (Berry, 1941; Esseen, 1942b, 1942a), éste es el caso si el tercer momento centrado $E((X_i - \mu)^3)$ existe y es finito. Entonces la velocidad de convergencia se mueve al menos con $\frac{1}{\sqrt{n}}$.

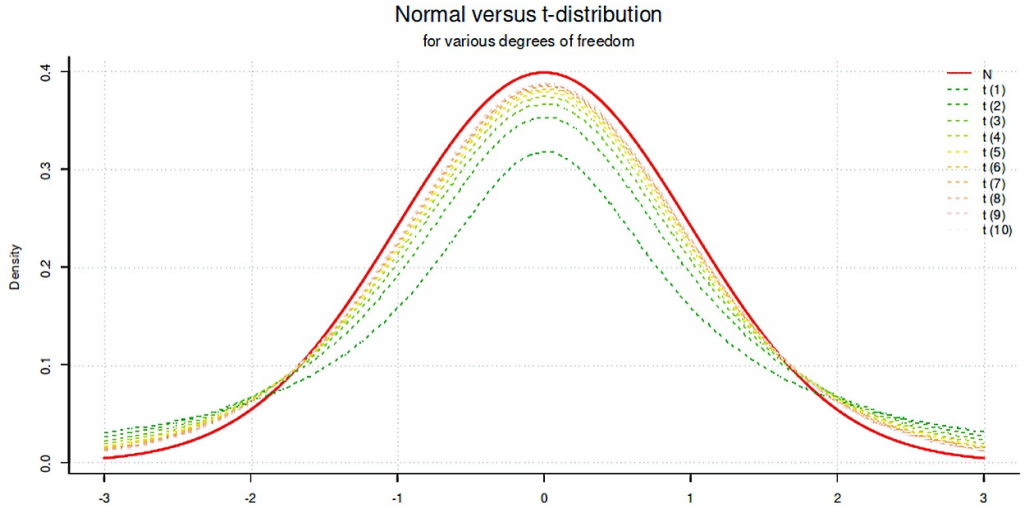


Figura 4.29. *Distribución normal frente a distribución t (diferentes grados de libertad)*

También es cierto que si existe una suma de variables aleatorias idénticamente distribuidas, su media (media aritmética) también está aproximadamente distribuida normalmente, ya que el paso de la suma a la media no es más que una transformación lineal de la forma

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad (4.19)$$

Y por lo tanto

$$\frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)} \sim N(0,1) \qquad (4.20)$$

Esta circunstancia es la base de la siguiente simulación. La función de R `clt.simulate()` simula funciones de distribución arbitrarias, calcula los valores medios y genera varios gráficos (véase la Fig.4.30): Histograma y estimación de la densidad para x y s, así como para un único conjunto de datos de la simulación y un gráfico de x frente a la distribución normal teórica. Las llamadas a las funciones aleatorias de la función de distribución de interés (por ejemplo, `rnorm(), rt(), rchisq()`, etc.) se colocan dentro de la llamada a `quote()`. Esta función permite que guardar cualquier expresión R para su posterior ejecución. El siguiente ejemplo genera números aleatorios de una distribución normal con $\mu = 4$ y $\sigma = 2$ y un tamaño de muestra de $n = 30$. Esto se repite 100 veces y el valor medio y la desviación estándar se escriben en un tabla (`ptII_quan_classicstats_centrallimittheorem.r`).

```
# define some samples as 'function to call'
fun1 <- quote(rnorm(n=30, mean=4, sd=2)) fun2 <- quote(rnorm(n=50, mean=4, sd=2)) fun3
<- quote(rnorm(n=100, mean=4, sd=2)) # simulation central limit theorem
clt.simulate(fun=fun1, trials=100) clt.simulate(fun=fun2, trials=10000)
clt.simulate(fun=fun3, trials=100)
```

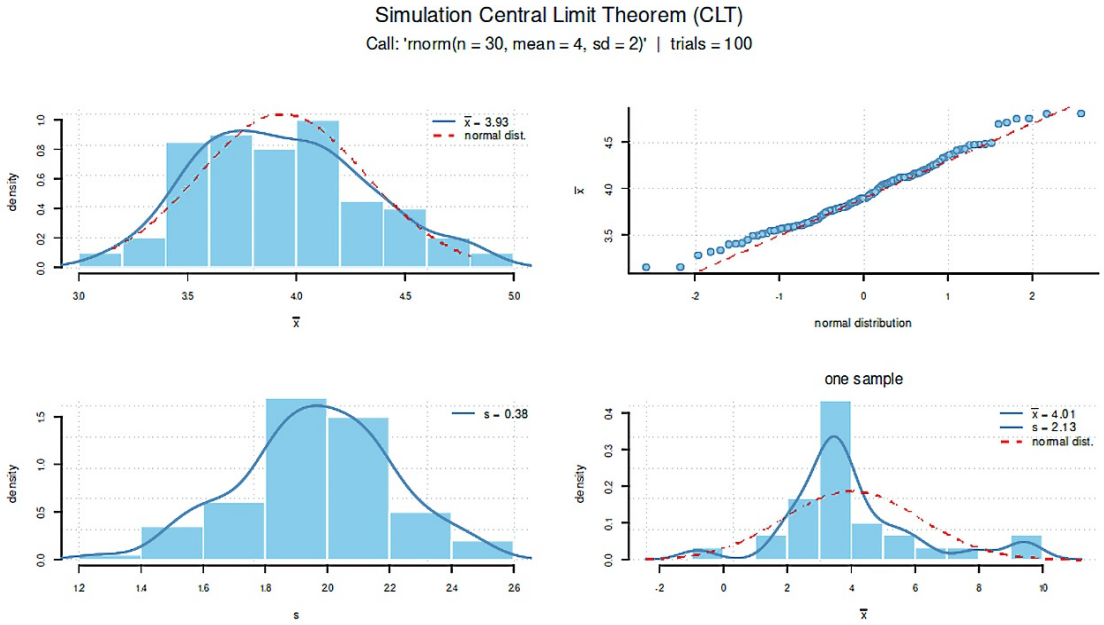


Figura 4.30. Simulación del Teorema del Límite Central
(valores aleatorios distribuidos normalmente, ensayos = 100)

Como puede verse en la Figura 4.30, con tan pocas réplicas y un tamaño de muestra relativamente pequeño, la aproximación a la distribución normal sigue estando en un rango modesto. Especialmente en los extremos, en parte no es cierto en absoluto cuando se compara la distribución empírica con la distribución normal teórica (arriba a la derecha en la Fig. 4.30).

Lo mismo puede verse al observar una sola simulación (parte inferior derecha de la Fig. 4.31). 10 000 repeticiones muestran una imagen claramente mejorada (véase la Fig. 4.31):

```
clt.simulate(fun=fun1, trials=10000)
```

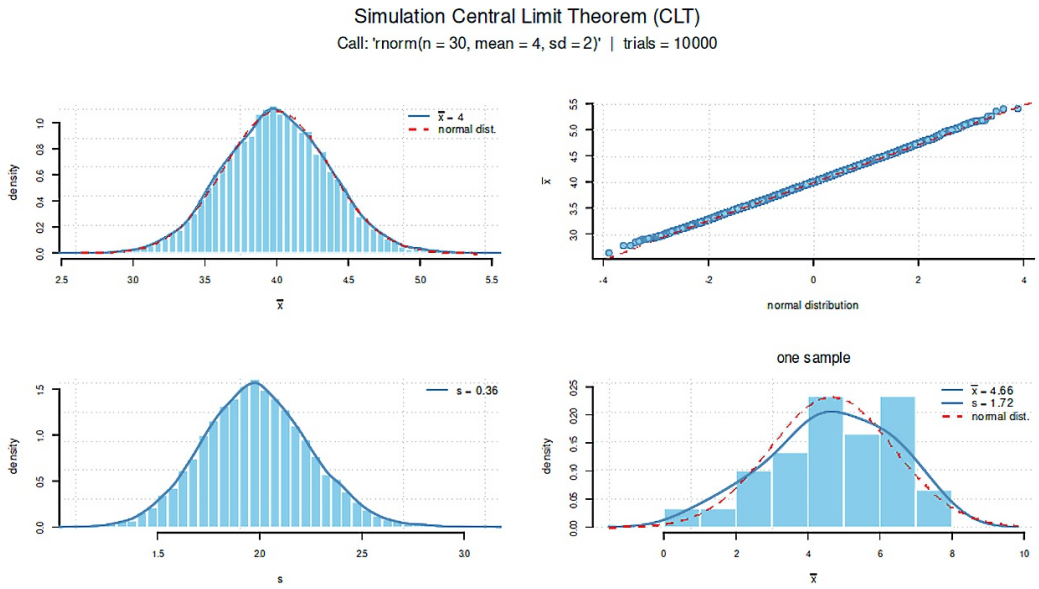


Figura 4.31. *Teorema del límite central de simulación (valores aleatorios distribuidos normalmente, ensayos = 10 000).*

Ahora la distribución normal simulada y la teórica para los valores medios apenas se distinguen, mientras que las distribuciones individuales (abajo a la derecha en la Fig. 4.31) siguen estando lejos de la distribución normal, como era de esperar. Que esto no sólo es válido para variables aleatorias con distribución normal – el núcleo del teorema del límite central – se demuestra con un simple cambio de la función de distribución aleatoria (véase la Fig. 4.32), si ahora se utiliza como base la distribución χ^2:

```
fun5 <- quote(rchisq(n=30, df=5, ncp=0))
clt.simulate(fun5, trials=10000)
```

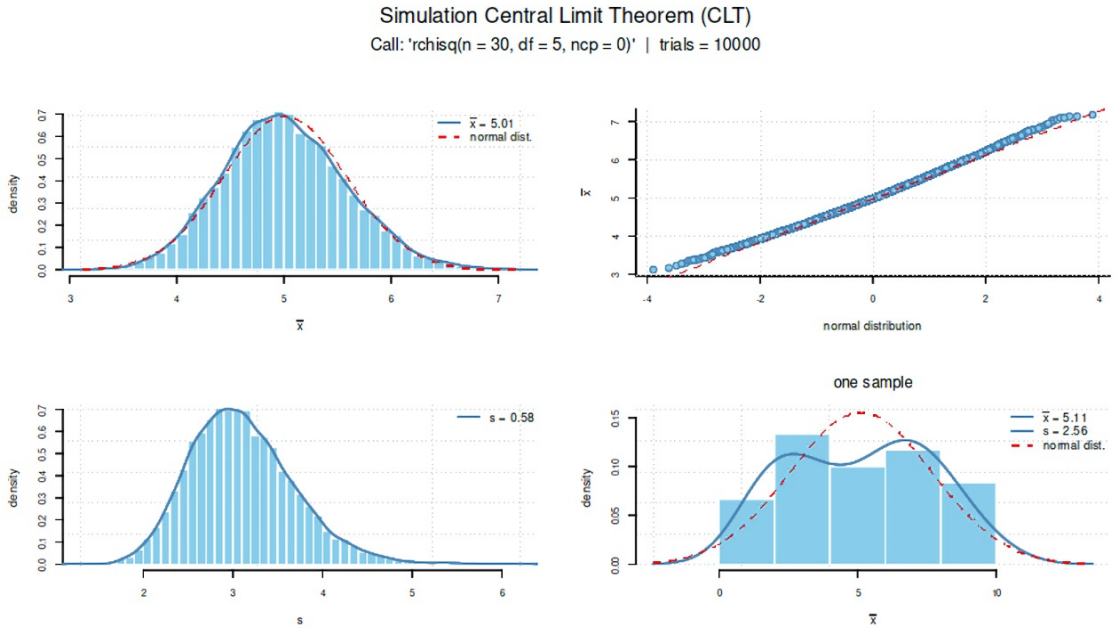


Figura 4.32. *Simulación – Teorema del límite central (valores aleatorios distribuidos en χ, ensayos = 10'000).*

La función `compare.clt.sim()` permite generar conjuntos de datos para diferentes tamaños de muestra y el mismo número de simulaciones:

```
Ns <- c(15, 30, 50, 100, 150, 1000, 10000)
trials <- 1000
seed <- 9182
res.clt.sim.Ns <- clt.sim.diffN(fun=fun6, Ns=Ns, norma=FALSE,
seed=seed, trials=trials)
str(res.clt.sim.Ns)
```

Para visualizar el resultado, se puede utilizar `clt.sim.diffN()` (véase la Fig. `4.33`):

```
res.clt.sim.Ns <- clt.sim.diffN(fun=quote(rnorm(n=X, mean=4, sd=2)),
           Ns=Ns, norma=FALSE, seed=123565, trials=pruebas)
str(res.clt.sim.Ns)
```

El parámetro `norma=FALSE` representa las distribuciones simuladas sin normalización, es decir, sin ajuste a x = 0 y s = 1. El cambio a `norma=TRUE` realiza la normalización a los coeficientes de la distribución normal estándar (véase la Fig. 4.34),

```
clt.sim.diffN(fun=fun6, Ns=Ns, norma=TRUE, seed=seed, trials=trials)
```

para que la aproximación a la distribución normal sea más clara. En cambio, las diferencias entre los tamaños de las muestras destacan mejor de otra manera en el primer ejemplo no normalizado.

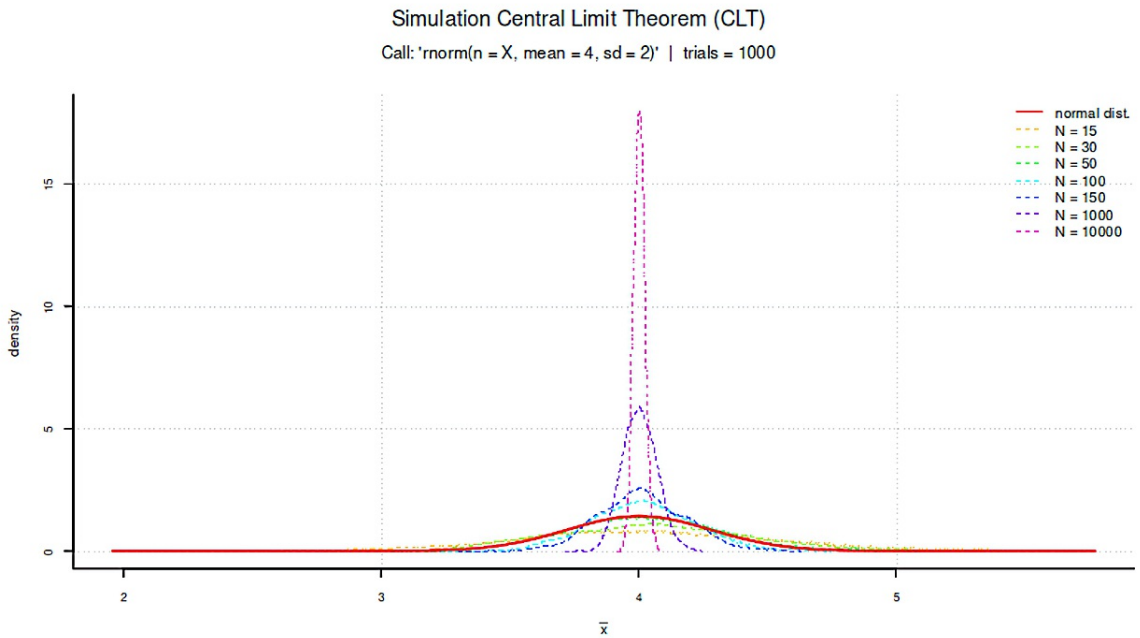


Figura 4.33. *Simulación – Teorema del límite central (Varios tamaños de muestra, no estandardizados)*

El hecho de que no se trata simplemente de aproximar valores arbitrarios a la distribución normal, sino de aproximar la distribución de los valores medios de una distribución, se muestra en la figura 4.35. Aquí, en la esquina superior derecha, están un sorteo arbitrario de valores distribuidos $\chi 2$, en la esquina superior izquierda la distribución media de valores distribuidos $\chi 2$, y en la esquina inferior izquierda la comparación de esta distribución media con la distribución normal. Es fácil ver que incluso con un tamaño de muestra de

N = 100, la distribución media de las extracciones aleatorias distribuidas en $\chi 2$ con 50 grados de libertad cada una consigue una muy buena aproximación a la distribución normal.

Sin embargo, el teorema del límite central per se no garantiza que, a partir de un determinado tamaño de muestra, los valores medidos se distribuyan realmente de forma normal. Por el contrario, el teorema central del límite trata de distribuciones muestrales y no de la distribución de los valores medidos. Dado que muchas cantidades suelen tener una distribución natural normal, puede surgir esta ilusión. No hay que olvidar que existen, por ejemplo, distribuciones bimodales y que incluso con tamaños de muestra muy grandes las distribuciones pueden seguir desviándose significativamente de una distribución normal (Wilcox, 1995). Por otro lado, la distribución *t*, por ejemplo, es relativamente robusta a las violaciones del requisito de distribución normal (Wilcox, 2012), de modo que los resultados basados en la distribución t son correspondientemente más robustos-válidos y, por tanto, más fiables en términos relativos. En la estadística clásica, el teorema del límite central se utiliza para justificar la realización de pruebas de hipótesis incluso cuando la población de referencia no está necesariamente distribuida de forma normal, pero la muestra de que se dispone es aproximadamente lo bastante grande.

Esta discusión pretende ilustrar que los datos distribuidos normalmente son el prerrequisito común en la estadística clásica, es decir, la distribución normal es el prerrequisito para tomar decisiones de prueba que se ajusten aproximadamente o, en otras palabras, que ni subestimen ni sobreestimen los efectos que se producen y se ajusten a los procedimientos de análisis de datos elegidos. Si los datos se desvían demasiado de los supuestos y si los procedimientos de análisis son sensibles y poco robustos frente al incumplimiento de sus requisitos previos, se tomarán decisiones equivocadas. Entonces se toma una decisión a favor de H_1, por ejemplo, aunque en realidad H_2 sería válida. Así, dependiendo de la situación de los datos y de la distribución subyacente, se introducen distorsiones y las tasas de error α o β aumentan desproporcionadamente en el sentido de la teoría de Neyman-Pearson. Surgen alternativas a lo largo de la reducción del nivel de datos y la elección de un procedimiento de análisis diferente. Esto conlleva el problema de cambiar la forma de la hipótesis y, en última instancia, las proposiciones. También es posible utilizar técnicas EDA (véase el capítulo 5) o la estadística de Bayes (véase el capítulo 6) o simplemente la simulación. También puede ser posible y útil revisar la selección y el tamaño de la muestra o incluso todo el diseño o los instrumentos de medición. La reducción del error de medición no suele tenerse suficientemente en cuenta en las estadísticas. Sea cual sea la opción elegida, el esfuerzo es considerable.

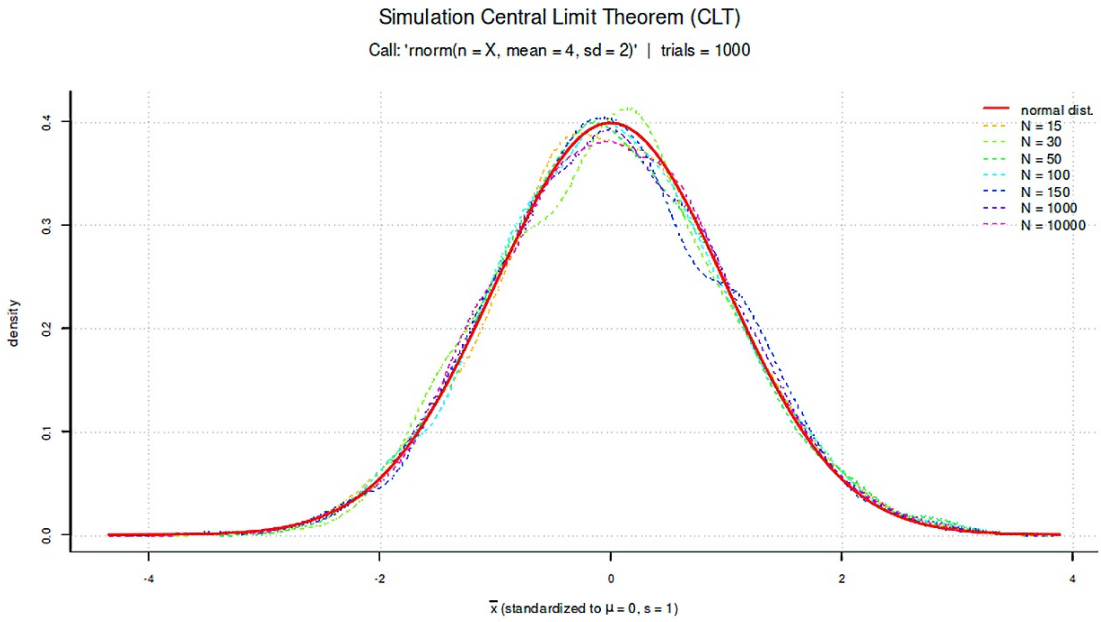


Figura 4.34. *Simulación – Teorema del límite central (Varios tamaños de muestra, estandardizados)*

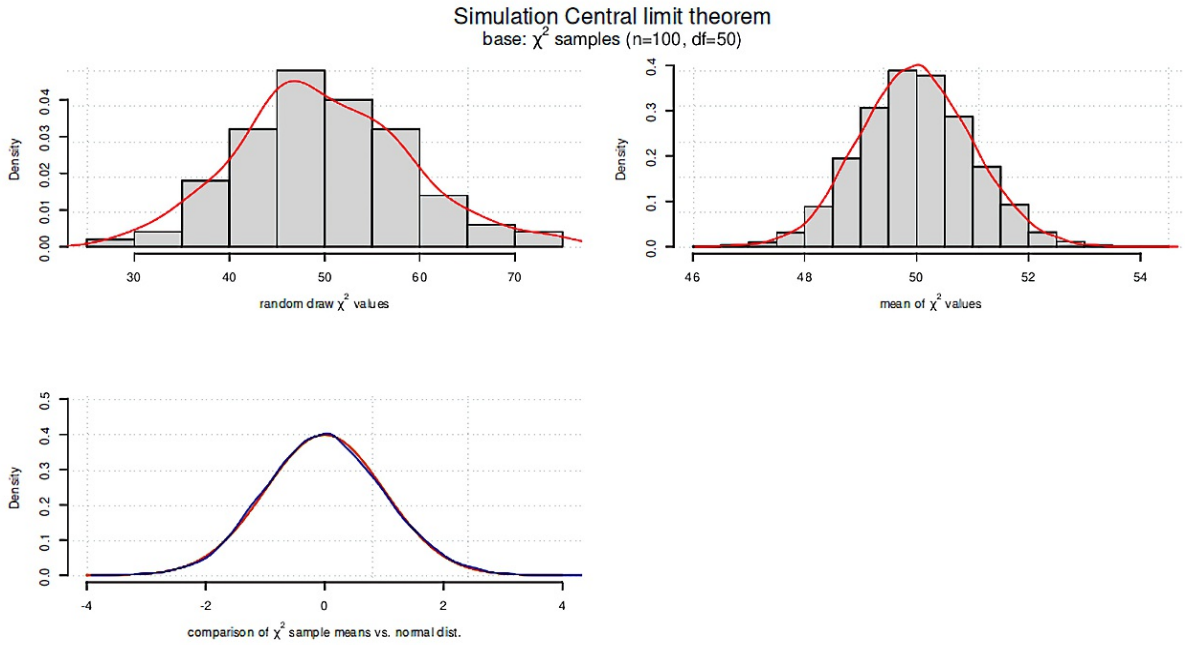


Figura 4.35. *Simulación – Teorema del límite central* (Aproximación por valores distribuidos de modo χ^2)

4.3.10 Epistemología recargada – estadística clásica

Desde el punto de vista de la teoría de la ciencia, se puede hacer un apunte: La estadística clásica no responde a preguntas de investigación relevantes porque no puede calcular cantidades numéricas – probabilidades – y, por tanto, no tiene respuestas. En cambio, da respuestas a preguntas que generalmente no se plantean, a saber, informando del valor *p*. Todo esto no se debe a las matemáticas subyacentes ni al procedimiento de análisis de datos elegido. Más bien, el problema se basa exclusivamente en una concepción epistemológica que supone que la probabilidad de los datos bajo la validez de una hipótesis nula más o menos específica puede decir algo sobre la probabilidad de una hipótesis dados los datos empíricos de que se dispone. Esto no es posible; por lo tanto, el problema no reside en los análisis per se, sino en el planteamiento básico, así como en la interpretación de los resultados y las conclusiones derivadas (Loredo, 1990, 1992). Por cierto, es posible un equivalente bayesiano para todos los análisis imaginables. Como se ha mostrado anteriormente, tanto la teoría de Fisher como la de Neyman-Pearson pueden muy bien utilizarse de forma epistemológicamente significativa. Este es el caso cuando Neyman-Pearson se utiliza en la garantía de calidad. Y Fisher se aplica en contextos ambiguos para encontrar sucesos estadísticamente raros y en realidad es relevante la probabilidad de los datos y no la de las hipótesis. La aplicación de todo esto conduciría entonces a una comprensión de la investigación fundamentalmente diferente. Sin embargo, dado que sostenemos que el conocimiento es relativo y debe provenir principalmente de la propia intuición, el problema epistemológico de la estadística clásica se formula de la siguiente manera como tarea para los lectores interesados:

Tarea 4.5: Interés de la investigación

Piense en una pregunta de investigación de su elección o tome una de la bibliografía o de su propio trabajo. Si en este caso se ha utilizado o debería utilizarse la estadística, ¿qué le interesa? ¿Por la probabilidad de los datos dados o más extremos bajo la validez de la hipótesis nula - o por la probabilidad de modelos (hipótesis) más o menos complejos? Discuta ambos casos, es decir, qué conclusiones obtiene en cada caso y cuáles no. Al hacerlo, refiérase directamente a los datos disponibles y a las hipótesis existentes y busque el área para la que se da información, el área para la que no y si y cuáles no, así como si las conclusiones se justifican en última instancia y cómo.

Lamentablemente, en casi todos los manuales de estadística falta otro punto: ¿cuál es exactamente la probabilidad de los datos que existen empíricamente y están disponibles? Si los datos están disponibles, se podría argumentar, por ejemplo, que los datos empíricos recogidos tienen una probabilidad de p = 1 y, por tanto, ya no tienen ninguna probabilidad, porque son precisamente estos datos los que se han manifestado y no otros. En este sentido, ya no tienen una probabilidad, sino que se han convertido en realidad. De lo contrario, habría que preguntarse desde qué perspectiva relativa se determina esta probabilidad, si los datos se han hecho reales y ya no representan un potencial. De todos los conjuntos de datos posibles, y hay muchos, a saber, una cantidad cuasi infinita (dependiendo del error de medición), se han manifestado exactamente éstos y ningún otro dato.

Esta cuestión ya ha sido debatida de forma modificada por Neyman-Pearson en relación con los intervalos de confianza. Se trata de un problema epistemológico y no matemático. También habría que preguntarse si los datos sólo tienen potencialmente una probabilidad antes de su realización, pero ya no desde su manifestación.

Hay una afirmación interesante en la teoría de Neyman-Pearson, que aborda un problema equivalente. Así, Neyman (1937) subraya con respecto a los intervalos de confianza si un parámetro medido real se encuentra o no en el intervalo de confianza. Sin embargo, no existe ninguna probabilidad de que esto ocurra. Por supuesto, podemos calcular la probabilidad de los datos cuando se trata de una predicción. Pero esa no es la cuestión aquí. Se trata de los propios datos empíricos. Tanto más sorprendente es entonces que una discusión sobre la cuestión de *si los datos manifiestos tienen una probabilidad de p = 1* ,no se lleve a cabo, sino que precisamente se calcule esta probabilidad bajo la validez de la hipótesis nula para decidir sobre los resultados de la investigación. Ahora se podría pasar a considerar la probabilidad de los datos antes de tamizarlos, pero las probabilidades a priori no se dan en la estadística clásica. La traducción real del valor p para responder a una hipótesis de investigación es entonces un acto cualitativo, no cuantitativo. Esto ya no se justifica en el propio procedimiento de análisis de datos, puesto que a partir de su interpretación, en sentido estricto, se abandona la estadística; al fin y al cabo, de lo que se trata es de obtener sentido y significado a partir de los números.

Parece mucho más interesante cuán probable es la hipótesis nula u otra hipótesis a la vista de los datos empíricos (= manifestaciones reales del objeto de investigación). Éste es precisamente el objetivo de la estadística de Bayes. Así, las hipótesis, que compiten entre sí, pueden contrastarse o integrarse en un modelo más complejo. Por ejemplo, a uno le gustaría saber si una hipótesis H_1 puede explicar los datos mejor que una hipótesis H_2 o H_3 o H_4, etc., y si entonces esto también se aplica a datos futuros. Este camino conduce directamente a los factores de Bayes. Otros autores, como el estadístico Andrew Gelman, abogan por un modelo complejo en lugar de dejar que los modelos que compiten entre sí se eliminen mutuamente y sigan quedando por debajo de las posibilidades porque hay muy poca integración. Esto hace que el debate pase de una lógica de "*o bien ... o ...*" a una lógica de "*tanto ... como*", que probablemente pueda ampliarse considerablemente con la lógica budista del tetralema.

Gelman persigue un enfoque fundamentalmente dialéctico de integración y síntesis, no para demostrar una supuesta superioridad, sino para preservar la complejidad sin una gran pérdida de información y así hacer más justicia a la realidad.

A modo de ejemplo, cabría preguntarse si el comportamiento humano a la hora de votar en tiempos de globalización y terror puede explicarse más bien por el miedo y la ira ("emociones"), por el razonamiento cognitivo ("elegir el mal menor"), el rechazo ("expulsar a esta persona"), por una combinación de todos estos factores o incluso por algo totalmente distinto. Todas estas afirmaciones podrían reformularse en hipótesis y contrastarse entre sí. Se podrían obtener probabilidades para cada una de estas hipótesis y, a continuación, seguir desarrollando el trasfondo teórico, predecir comportamientos u ofrecer intervenciones significativas. Sin embargo, sería más útil, como sugiere Gelman, preguntarse qué combinación de posibles factores de influencia puede explicar la mayor parte de los datos, así como los casos especiales y los extremos. La complejidad del modelo parece mucho más importante que las probabilidades de hipótesis singulares. Un experimento mental sobre exactamente esto hace hincapié en el hecho que parece muy poco probable que todas las hipótesis tengan exactamente la misma probabilidad. Según el instrumento de medida, el método de encuesta, etc., una u otra hipótesis recibirán una probabilidad relativa mayor que las demás y podrían así *prevalecer*. Sin embargo, como las demás también tienen una probabilidad legítima queda rápidamente claro que no se trata de eso. Decidirse por una sola hipótesis significaría sencillamente desechar explicaciones alternativas legítimas y, como tales, información valiosa. Eso es ineficaz e imprudente. En este sentido, no es necesario realizar cálculos numéricos para adoptar una postura fundamental que favorezca los modelos integradores complejos. Sin embargo, hay excepciones, es decir, siempre es necesario analizar si tiene sentido un modelo simple o un modelo complejo.

Corresponde a cada uno cual "creer", si la rareza de los datos es relevante para determinar la validez de las hipótesis. Existen ciertas soluciones (work-arounds) que permiten determinar los factores de Bayes a partir de los valores *p* mediante la llamada *calibración p* (Sellke, Bayarri & Berger, 2001) y, de este modo, contrastar las hipótesis entre sí de forma indirecta. Los factores de Bayes se derivan entonces de una combinación de valor *p* y *probabilidad a priori* añadida (véase también el capítulo 6.8.1). En consecuencia, los factores de Bayes presentan todos los problemas que ya caracterizan a los valores *p*. Posiblemente, también los que pueden surgir mediante la adición de probabilidades a priori (Schimmack, 2015b). Por lo tanto, es cuestionable que esto prometa una ganancia real de conocimiento, excepto, por supuesto, por la posibilidad de incorporar conocimiento previo a los p-valores. El procedimiento no es inequívoco, ya que la elección de las distribuciones previas, por ejemplo, no parece surgir necesariamente de una comprensión cualitativa previa del contexto específico, sino que fue seleccionada, al menos en parte, sobre una base matemático-pragmática o incluso por convención, véase, por ejemplo, la discusión en Schimmack (ibid.). Los factores de Bayes tampoco corresponden a un análisis bayesiano completo. Sin embargo, mientras la base matemática no esté anclada en el objeto y el contexto de la investigación, representa una decisión digna de discusión o incluso cuestionable. Entonces sería mejor examinar la ganancia de conocimiento bajo diferentes priors, así como con y sin el uso de factores de Bayes, con el fin de examinar los resultados y su estabilidad más de cerca. El resultado final es de nuevo un tipo de análisis de diseño como el descrito por Gelman y Carlin (2014). De esto aprendemos que uno debe examinar cuidadosamente el diseño, ya sea clásico o bayesiano, en términos de la magnitud y la dirección de los efectos encontrados.

> Tarea 4.6: Criterios y decisión
>
> ¿Basado en qué criterio sustantivo se toma una decisión? ¿Cómo lo han puesto en práctica los lectores en sus propias investigaciones hasta el momento?

Utilizando el sentido común, cabe preguntarse: si una magnitud como el *valor p* es convertible más o menos directamente en otra como el *factor de Bayes* de forma bastante inequívoca, y viceversa, ¿sigue existiendo realmente una diferencia cualitativa entre estas magnitudes y cómo se concreta? En el caso que nos ocupa, esta diferencia consiste simplemente en la adición de información a priori, sin por ello llevar a cabo un análisis plenamente bayesiano. También cabe preguntarse si, en cualquier circunstancia, un análisis

frecuentista debe doblarse de tal forma que se obtenga un resultado "medio" bayesiano. Esto puede ser útil en los metaanálisis cuando sólo se dispone de resultados publicados, como los valores *p* o *t*. Parece menos útil cuando se dispone de datos brutos y se planifica y realiza un estudio. En este caso, es obvio elegir inmediatamente una implementación bayesiana completa, que, sin embargo, no requiere esencialmente factores de Bayes.

Aparte de eso, la discusión hasta ahora no responde a la cuestión relevante de la significación y el razonamiento que subyace a las decisiones, es decir, cómo justificar las decisiones en términos de contenido y en el contexto de los datos. Esta cuestión es independiente de la bayesiana, la clásica o incluso de la cuantitativa y la cualitativa. Como dice Gigerenzer (2004b, p.590f.) sobre el procedimiento de la teoría de Neyman-Pearson, el diseño consistente en H_1, H_2 así como α, β, N y se le crea antes del experimento "basándose en consideraciones subjetivas de coste-beneficio". Las decisiones están sujetas a la discreción subjetiva del investigador y, como tales, requieren una justificación – otra vez más una circunstancia que requiere teóricamente un anclaje en la materia, pero que a nivel práctico es muy difícil de justificar. Tanto si se elige un nivel de significación como criterio de decisión o una probabilidad dados los factores de Bayes. En cada caso, debe estar fundamentado en la materia. Si no lo está o no puede estarlo, falta un argumento importante de la cadena de razonamiento científico-lógico e – independientemente de si es bayesiano o clásico – los representantes deben aceptar el reproche de arbitrariedad y de elección insuficientemente justificada de las barreras de significación. Además, como señalan Gelman y Stern (2006), la diferencia entre significativo y no significativo no es en sí misma significativa. Esta lógica también puede aplicarse a los factores de Bayes. Así pues, no sólo existen similitudes estructurales entre los métodos clásico y bayesiano.

Por supuesto, hay que tomar decisiones en algún momento y en algún lugar. Nadie pretende que esto sea trivial. De lo que se trata aquí es simplemente de no adoptarlas qua convenciones y sin reflexión, sino de adaptarlas, sin duda laboriosamente, pero de forma adecuada a cada caso – incluyendo la anticipación de las consecuencias (incluso para la elección de la distribución y los p-valores resultantes). Los estudios de casos sobre Neyman-Pearson (véase el capítulo 4.3.3) ofrecen material suficiente para el pensamiento independiente y para recorrer escenarios de investigación y práctica. Esto lleva a una tarea para los lectores (véase abajo, tarea 4.7: "Decidir y justificar").

En la estadística frecuentista es legítimo formular una hipótesis nula en el nivel crítico (1 %, 5 %, ... según la convención). Sin embargo, una hipótesis nula en sí misma no recibe una probabilidad confirmatoria. Por tanto, *no puede aceptarse*, sino simplemente *no rechazada o retenida*. En cuanto al procedimiento y la elección del lenguaje, esto no refleja más que el procedimiento falsificador de Popper y no representa el espectro del conocimiento científico-teórico legítimo. Así pues, sigue sin estar claro en qué consiste realmente la ganancia de conocimiento *al retener una hipótesis nula*. El *no rechazo* no genera ninguna ganancia de conocimiento, ya que sólo significa que los datos disponibles son compatibles con la hipótesis nula y no puede rechazarse al nivel crítico seleccionado. Aún así, son posibles muchas otras hipótesis que pueden conducir al mismo resultado – mantener H_0.

> Tarea 4.7: Decidir y justificar
>
> Crea tu propio escenario de obstáculos a la decisión y justificaciones. Piense en un caso práctico sencillo que trate explícitamente de la toma de decisiones y defina criterios de decisión concretos. criterios de decisión concretos. A continuación, tradúzcalos al pensamiento estadístico clásico y y genere datos mediante simulación R. A continuación, examine los posibles resultados, compárelos con los datos y saque conclusiones basadas en los números. Considera ahora qué ocurriría si cambiaras los criterios de decisión que elegiste al principio. cambiados. ¿Qué justificaciones podrías dar a los terceros por qué y cómo concretamente estos cambios tendrían sentido? Piense en los criterios de decisión cuando de repente queda el área sensata y justificable.
> ¿Qué cambiará en términos de contenido, qué cambiará estadísticamente?

En contextos de investigación y lejos de los entornos industriales, es difícil o incluso imposible controlar el tamaño de las muestras y el tamaño de los efectos. En el contexto de un cálculo de costes completo (es

decir, un análisis coste-beneficio), las tasas de α- y β-errores no pueden ajustarse de forma óptima, lo que en última instancia conduce a una subestimación o sobreestimación de los efectos. En la práctica, sin embargo, esto a menudo significa demasiados falsos positivos que no se pueden replicar (Szucs & Ioannidis, 2017-08-03). Una de las razones podría ser que la ciencia trata principalmente sobre el conocimiento de la novedad, lo que se refleja tanto en las prácticas de investigación como en las políticas de publicación de las revistas pertinentes. Este suele ser muy modesto o incluso inexistente cuando no se rechaza una H_0. Esto se vuelve problemático cuando lo que está en juego es precisamente la prueba de identidad o equivalencia (véase el capítulo 4.4.9). Un ejemplo sería el sector farmacéutico, donde un nuevo medicamento debe demostrar que tiene tan pocos (o tantos) efectos secundarios como un medicamento ya comercializado, por ejemplo, de modo que no sea posible una desviación a la baja. Una solución a nivel numérico es realizar las llamadas pruebas de equivalencia. Éstas se basan en definir primero un rango de diferencias de cero (caso ideal) u otro tamaño adecuado en términos de contenido que marque el rango dentro del cual se puede determinar que algo no presenta diferencias. Ahora se plantean dos hipótesis, pero formuladas de tal manera que ahora hay dos H_0. Éstas son: La diferencia de las cantidades investigadas se encuentra por encima o por debajo del intervalo de diferencia especificado. La H_1, a su vez, se centra directamente en el intervalo de diferencia. En caso de rechazo de ambas H_0 en el nivel dado, se considera que la equivalencia está asegurada. Otra posibilidad es calcular intervalos de confianza clásicos y comprobar si los valores que se legitiman como equivalencia se encuentran dentro de este intervalo, lo que a su vez indica que existe equivalencia (Dixon, Saint-Maurice, Kim, Hibbing, Bai & Welk, 2018).

> Tarea 4.8: Cuestionar los libros de texto
>
> Se invita al lector interesado a comprobar por sí mismo las tesis mencionadas. Investigue por su cuenta en los libros de texto que utiliza. ¿Qué informan y enseñan y ¿qué no abordan?

Teniendo en cuenta estos puntos de argumentación, cómo se hace hoy en día el razonamiento estadístico clásico, el enfoque original de Neyman-Pearson parece muy auténtico. Neyman-Pearson no trata de conocimientos, sino de decisiones basadas en datos dentro de unas condiciones marco predefinidas. El nivel específico del valor *p* no importa, sólo la comparación con el umbral crítico en el contexto de un análisis de potencia a priori. El grado de veracidad o falsedad de una hipótesis no puede deducirse de ello y no tiene mayor interés. El nivel de error en sí sólo se utiliza para tomar una decisión, y se basa en los datos. No tiene un fin en sí mismo. De este modo, Neyman-Pearson difiere de Fisher, que se ocupa de la inferencia y el conocimiento, no de la toma de decisiones a ciegas.

Sigue siendo incomprensible por qué esta discusión – que puede encontrarse fácilmente en muchos libros sobre estadística bayesiana (Jaynes, 2003; Gelman, Carlin, Stern y Rubin, 2004; Kruschke, 2015b)- nunca se encuentra en los libros de texto sobre estadística frecuentista de esta forma detallada o reducida. Sin embargo, puede encontrarse en artículos de revistas de estadística clásica (por ejemplo, Brandstätter, 1999) y en los de estadística bayesiana de todos modos.

La significación en sí misma suele tener el aura de ser la única fuente de conocimiento. La significación estadística – o no – se determina sobre la base de un exceso de probabilidad crítica que debe especificarse en términos de contenido en comparación con el valor *p* empírico. No añade ninguna información nueva, ya que no se basa en datos (información), sino en expectativas formuladas. Si cambian las expectativas, cambia la significación – y esto con datos idénticos. Podemos así afirmar:

Recordatorio 4.3: Significación estadística

La significación estadística la determinan los seres humanos. No es tarea de la estadística ni de los algoritmos, ni siquiera de los datos empíricos.

Históricamente, la significación alcanzó su posición dominante como convención general gracias a la presión ejercida hasta mediados del siglo XX. El propio Fisher es en gran parte responsable de ello y de su tendencia a producirse a sí mismo a través del cabildeo (Jaynes, 2003, cap. 16). Esto llegó tan lejos que muchas revistas no publicarían artículos que no pudieran demostrar los más altos niveles de significación sin entender cómo se produce la significación en primer lugar. En general, el lema era y sigue siendo: cuanto más pequeño, mejor. Sin embargo, ya en la década de 1990 la APA (1996-03, 1996-12-15), siguiendo las recomendaciones del tardio Fisher (1956/1973), trató de desplazar la atención hacia los valores *p* exactos y reducir la relevancia de la significación o la prueba de hipótesis nulas, pero sin desterrar estos conceptos (Wilkinson y el Grupo de Trabajo de la APA , 1999; Fidler, 2010). En su lugar, se requieren intervalos de confianza, tamaños del efecto y cálculos de potencia. Esto marcó básicamente un lento giro desde las pruebas hacia una mayor estimación.

Durante décadas, diversos autores (Cohen, 1962) han reclamado la comunicación de los tamaños del efecto, el cálculo a priori de los análisis de potencia, etc. Sin embargo, Sedlmeier y Gigerenzer (1989) pudieron demostrar en un meta-estudio que la potencia de los estudios relevantes publicados incluso disminuía con el tiempo debido a α-ajustes (¡!) – 24 años después del conocido estudio de Cohen (1962), que por primera vez criticó la falta de potencia de los estudios.

Por el contrario, en 2015 la revista "Basic and Applied Social Psychology" decidió prohibir por completo los valores *p* y las declaraciones de significación en sus publicaciones (Tramow & Marks, 2015; Wasserstein, 2015). De forma equivalente, Andrew Gelman anunció junto con su coautora Jennifer Hill (2007) prescindir en la próxima edición de su libro fundamental sobre los modelos lineales (jerárquicos) de cualquier referencia a los valores *p* y a la significación. La ASA (American Statistical Association) celebró un debate político sobre los valores *p* y la significación en 2014 (Wasserstein & Lazar, 2016; Greenland, Senn, Rothman, Carlin, Poole, Goodman & Altman, 2016, pero véase Ionides, Giessing, Ritov & Page, 2017). Esto suscitó nuevos debates (Matthews, Wasserstein & Spiegelhalter, 2017). El profesor de estadística estadounidense Georg Cobb resume la situación de la siguiente manera (Wasserstein & Lazar, 2016, p. 129):

> „Q: Why do so many colleges and grad schools teach p = 0.05?
> A: Because that's still what the scientific community and journal editors use.
> Q: Why do so many people still use p = 0.05?
> A: Because that's what they were taught in college or grad school."

Esto es similar a un argumento ofrecido por un adicto: la sustancia adictiva se ofrece como su propia solución (Wallace, 1998, p.38, cursiva en el original).

> „But something is *malignantly* addictive if (1) it causes real problems for the addict, and
> (2) it offers itself as a relief from the very problems it causes."

Es lógico que la ASA celebró una reunión del 11 al 13 de octubre de 2017 en Bethesda, MD, EE. UU., titulada "ASA Symposium on Statistical Inference. Scientific Method for the 21st Century: A World Beyond $p < 0.05$ (Simposio de la ASA sobre Inferencia Estadística. Método Científico para el Siglo XXI: un Mundo más allá de $p < 0,05$)" y publicó un número especial de acceso abierto sobre el tema como seguimiento de esta reunión. Otras revistas, como Nature Human Behavior (Benjamin et al., 2018-01-01) no hicieron una declaración clara sobre NHST, pero en cambio publicaron un comentario de 72(!!!) investigadores que, en el caso de NHST y nuevos descubrimientos, utilizaron un valor *p* reducido al nivel de convención habitual del 5% de $p < 0.005$ como criterio propuesto. En su pobre trabajo de justificación, los propios autores

discrepan entre sí. Así, por un lado, se señala dos veces que el criterio propuesto – al fin y al cabo, la afirmación central del artículo – es *completamente arbitrario* (ibid., p.7, p.8); ya Fisher lo rechazó (ibid., 182 Capítulo 4. Estadística clásica, p.8) y los autores entre sí prefieren en realidad otros métodos de análisis de datos (ibid., p.6, p.8) para presentar resultados y, por tanto, no apoyan en absoluto la NHST. Tanto más asombra tal comentario. Además, faltan las pruebas teóricas y empíricas alegadas de que el endurecimiento de la convención p al 0.5 % mejoraría la calidad de la investigación. Es de esperar lo contrario (palabras clave: p-hacking, sesgo de publicación, etc.). Los argumentos de la literatura sobre NHST y sus implicaciones para la investigación (por ejemplo, Meehl, 1967, 1990; Jaynes, 2003; Hubbard & Lindsay, 2008; Goodman, 1999, 2016; Colquhoun, 2014, 2017, 2018; y muchos más) apenas se tienen en cuenta en el artículo. Para empeorar las cosas es que se da la impresión de que simplemente se está cambiando el límite del *valor p* por un límite del *factor de Bayes*, como suele ocurrir con los cambios de poder sobre la interpretación de la realidad. Sin embargo, esto deja todos los problemas de los *p*-valores, que no se eliminan con la calibración de los *p*-valores (véase el capítulo 6.8.1.4). La fijación en un umbral arbitrario para separar la significación de la insignificancia también indica que presumiblemente no se comprende el continuo incierto de la investigación, por lo que deben crearse límites arbitrarios de significación libres de contexto. Por el contrario, hay suficientes marcadores, aparte de una sola estadística, para clasificar una investigación como digna de publicación. La atención en factores de Bayes fijos y umbrales críticos asociados no ayuda aquí (Gigerenzer & Marewski, 2015). En consecuencia, cada argumento de este artículo débil y casi vergonzoso es desmontado por una respuesta de McShane, Gal, Gelman, Robert, y Tackett (2019), mostrando que, por el contrario, de lo que se trata es de abolir por completo la noción de estadística de significación y los umbrales asociados. El valor *p* es entendido por los autores como el continuo incierto del que surge – entre otros factores a considerar en cuanto a la calidad de un trabajo de investigación. Los autores sólo dan deliberadamente pistas formuladas abiertamente sobre cómo se puede aumentar la calidad de los trabajos de investigación y organizar el trabajo de los revisores según criterios cuidadosos. Llegan a la conclusión de que no se trata de fijarse en un determinado valor numérico que siga siendo válido al final, independientemente del contexto, el ámbito temático, los trabajos anteriores, etc.

En cuanto a los valores p, no es cierto que "más pequeño" sea siempre mejor. La teoría estadística de Neyman-Pearson ya muestra con bastante precisión que tanto el α- como el β-índice de error sólo adquieren su significado en el contexto de todos los parámetros (fuerza de la prueba, tamaño de la muestra, tamaño del efecto, ...) y no tienen sentido por sí solos. Las simulaciones (por ejemplo, en el capítulo 4.3.5) pueden mostrar con qué facilidad pueden surgir valores *p* extremos, especialmente con muestras pequeñas – lo que es habitual en psicología – que desaparecen inmediatamente con tamaños de muestra mayores.

El tamaño del valor *p* es una función directa del tamaño de la muestra cuando el efecto está presente y las condiciones son constantes, y cuanto mayor sea *N*, menor será el valor *p*. En el caso de tamaños de muestra iguales, existe una relación directa entre los valores *p* y los tamaños de los efectos. Sin embargo, cuando éstos varían entre los estudios o dentro de ellos, los valores *p* conducen fácilmente a conclusiones engañosas (Brandstätter, 1999, p.10). Los tamaños de muestra grandes conducen a potencias de prueba máximas y viceversa. Con una fuerza de prueba máxima, las desviaciones mínimas se hacen visibles, independientemente de si son relevantes en la práctica o no. Tampoco existe una conexión directa con la existencia de una teoría científicamente sostenible en la que basar hipótesis derivables y comprobables. No hay que olvidar, y de nuevo, que el valor *p* depende en gran medida de la muestra.

El siguiente script R muestra (véase la Fig. 4.37) cómo, con un tamaño del efecto constante (diferencia media estandarizada *d* según Cohen, 1992) entre dos poblaciones conocidas distribuidas normalmente, los valores *p* tienden inevitablemente hacia cero (prueba *t*) sólo debido al aumento del tamaño de la muestra – y eso aumenta la ilusión de significación. El efecto práctico real (diferencia de medias en la en la población), por otra parte, permanece constante a pesar del aumento de la muestra y puede ser completamente insignificante en la práctica. Matemáticamente, resulta aún más claro, ya que el error típico – aquí para la media aritmética – sigue la ecuación

$$\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}} \tag{4.21}$$

y para N hacia infinito el error típico tiende a cero. Dado que los valores *t* (por ejemplo con coeficientes de regresión) pueden deducirse de la fórmula

$$t = \frac{\text{Parameter}}{SE_{\text{Parameter}}} \tag{4.22}$$

se puede ver que el aumento de N da como resultado valores *t* cada vez mayores y valores *p* cada vez menores. La siguiente conversión de valores *t* en valores *p* lo demuestra; se toman valores aleatorios con $\mu = 0.5$ y $\sigma = 1.2$ y se prueban frente a cero según NHST (`ptII_quan_classicstats_p-t-df-relationship.r`):

```
# seed <- 567
N <- seq(30,1000,10)
ts <- vector()
pvs <- vector()
mu <- 0.5
sd <- 1.2
usesameseed <- TRUE #FALSE
for(i in 1:length(N))
{
print(N[i])
if(usesameseed) set.seed(seed)
samp <- rnorm(N[i], mu, sd)
samp.t.test <- t.test(samp)
ts[i] <- samp.t.test$statistic
pvs[i] <- samp.t.test$p.value
}
par(mfrow=c(2,2), oma=c(2,1,4,1), "cex.axis"=1, bty="l")
plot(N, ts, bty="n", type="l", col="darkred", pre.plot=grid())
plot(N, log(pvs), bty="n", type="l", col="darkred", pre.plot=grid())
plot(ts, log(pvs), bty="n", type="l", col="darkred", pre.plot=grid())
mtext("Dependence N, t and p-value", side=3,
            line=0.8, cex=2, outer=TRUE)
mtext(paste("same seed = ",usesameseed,sep=""), side=3,
            line=-1.8, cex=1.5, outer=TRUE)
```

A medida que aumentan los valores *t*, disminuyen los valores *p* y el tamaño de la muestra *N* crece de forma constante. La relación entre los valores *t* y *p* es casi lineal, en la figura por *log(p)* se trata de una curva descendente. Con el mismo valor inicial del generador aleatorio, la influencia del tamaño de la muestra se hace evidente (arriba en la Fig. 4.36). Con otros valores iniciales, cabe esperar fluctuaciones debidas a la aleatoriedad, que pueden enmascarar inicialmente la influencia del tamaño de la muestra (parte inferior de la Fig. 4.36).

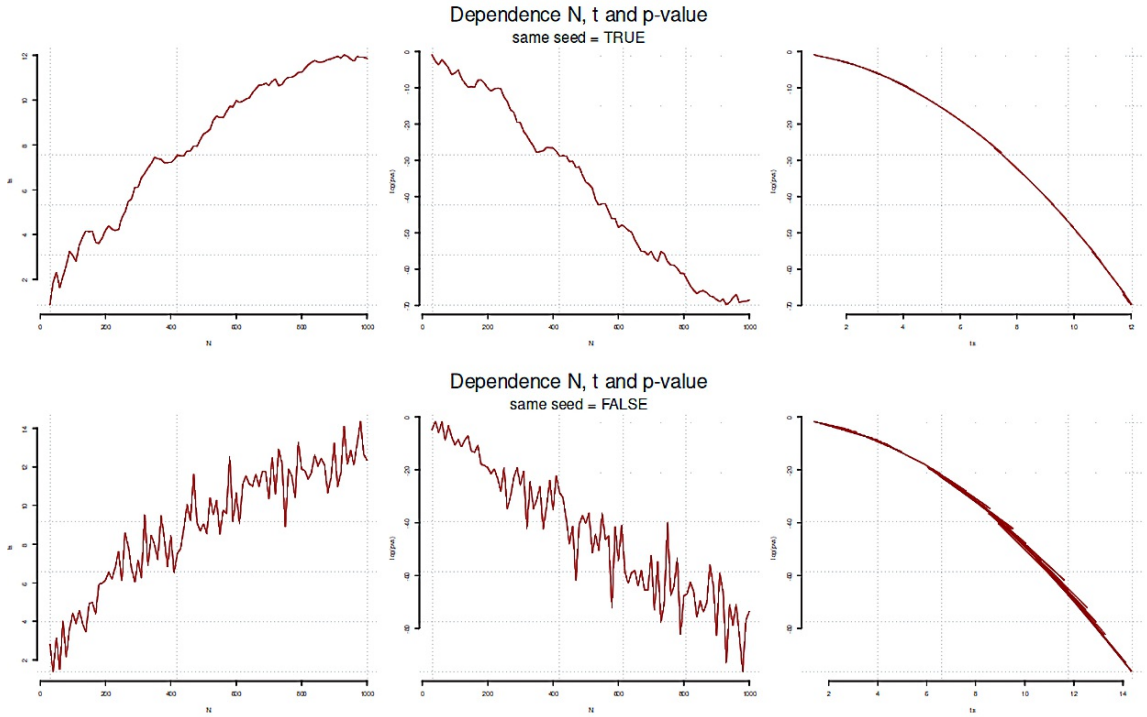


Figura 4.36. *Relación entre los valores N, t y p*
(arriba con el mismo, abajo con distinto valor inicial del generador aleatorio)

A su vez, la Fig. 4.37 muestra ahora la relación de los grados de libertad (= df) y los *p*-valores, cuando la estimación del parámetro (por ejemplo, el valor *t* de un coeficiente de regresión) se mantiene constante. En la escala logarítmica, a su vez, resulta aún más claro cómo la curva de los valores *p* disminuye constantemente al aumentar los grados de libertad y, por tanto, el tamaño de la muestra *N* (Fig. 4.37, `ptII_quan_classicstats_p-t-df-relationship.r`):

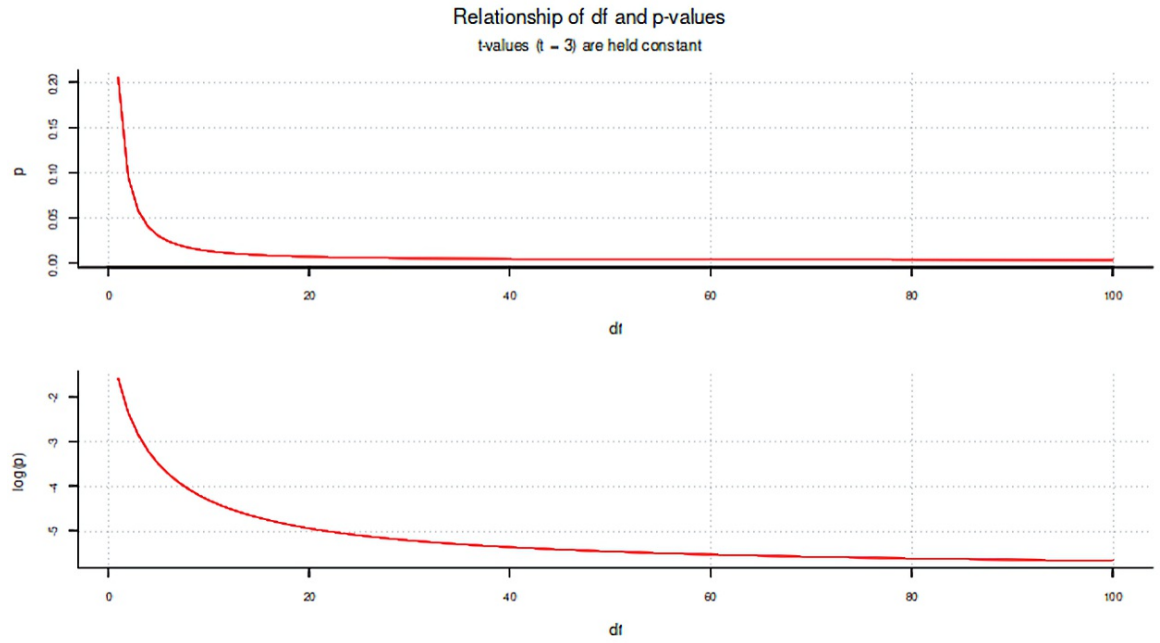


Figura 4.37. *Relación entre los valores p y los grados de libertad (valor t constante)*

```
# relationship df and p value in face of growing df's and const. t values
sek3 <- 1:100
tvalue <- 3
par(mar=c(4,4,2,1), oma=c(1,1,3,1), "cex.axis"=0.8, mfrow=c(2,1))
plot(sek3,2*pt(-tvalue,df=sek3), panel.first=grid(), type="l",
           col="red", bty="l", main="", xlab="df", ylab="p")
plot(sek3,log(2*pt(-tvalue,df=sek3)), panel.first=grid(), type="l",
           col="red", bty="l", main="", xlab="df", ylab="log(p)")
mtext(paste("Relationship of df and p-values",sep=""), 3,
           outer=TRUE, line=0.5, cex=1.4)
mtext(paste("t-values (t = ",tvalue,") are held constant",sep=""), 3,
           outer=TRUE, line=-1, cex=1.1)
```

Un análisis de potencia a priori – equivalente a la decisión de trabajar según Neyman-Pearson – no suele encontrarse en las publicaciones. En el mejor de los casos, existen análisis de potencia posthoc, que, sin embargo, sólo ponen en otras palabras el resultado empírico ya conocido y, por lo tanto, no prometen ninguna ganancia adicional de conocimiento, a menos que se lleve a cabo el análisis de diseño ya discutido (véase el capítulo 4.3.3.2). Una excepción es el uso de un análisis de potencia post-hoc para preparar una futura replicación con el fin de planificar cuidadosamente los tamaños del efecto, el nivel de significación objetivo y el tamaño de la muestra. Pero entonces esto se transforma en un análisis de potencia a priori y deja de ser un análisis de potencia post-hoc en sentido estricto. Se vuelve más interesante cuando los análisis se examinan mediante simulación bootstrap (paramétrica) (véase también Excursus sobre Simulación, Capítulo 4.3.5). Aquí es posible, además de una distribución de los parámetros por cada ejecución bootstrap y generar así una distribución de los valores *p*. El número de bootstraps para los que el valor *p* es menor que la probabilidad crítica de superación dividido por el número de todos los bootstraps realizados da como resultado el valor de potencia simulado. El siguiente código R con la función `sim.DiM()` demuestra esto con una simple prueba *t* de dos muestras independientes, cada una con un tamaño de muestra de $n_1 = n_2 = 25$ y $\sigma = 2$. Los valores medios difieren con $\mu_1 = 6.5$ y $\mu_2 = 7.8$. El resultado es una diferencia teórica de $\mu_1 = \mu_2 = -1.3$ y una d de Cohen de $d = (6.5 - 7.8)/2 = -0.65$.

```
# post hoc power via simulation
# bootstrap parametric model post hoc power
#
# t-test
seed <- 9876
set.seed(seed)
alpha <- 0.01
trials <- 1000
LM <- FALSE
#LM <- TRUE
sim.DiM <- function(LM=TRUE,n1=25, n2=25,mu1=6.5,mu2=7.8,s1=2,s2=2)
{
a <- round(rnorm(n=n1, mean=mu1, sd=s1))
b <- round(rnorm(n=n2, mean=mu2, sd=s2))
if(LM)
{
res <- summary(lm(a ~ b))$coef["b","Pr(>|t|)"]
} else
{
res <- t.test(a, b)$p.value
}
return(res)
}
power.boot.res <- replicate(trials, sim.DiM(LM=LM))
#describes(power.boot.res)
#summary(power.boot.res)
fivenum.wn(power.boot.res)
sd(power.boot.res)
# plot
```

```
par(mar=c(4,4,2,1), oma=c(1,1,3,1), "cex.axis"=0.8, mfrow=c(2,1))
hist(power.boot.res, panel.first=grid(), prob=TRUE, border="white",
           col="skyblue", main="", xlab="p", ylab="density", breaks="FD")
# restrict p-value due to density estimation algorithm
# only densities above zero and below 1
dens.pbr <- density(power.boot.res)
dens.pbr$x[dens.pbr$x <= 0] <- .Machine$double.eps
dens.pbr$x[dens.pbr$x > 1] <- 1-.Machine$double.eps
lines(dens.pbr, col="red", lwd=2)
# power for different p-values
alphas <- c(0.1,0.05,0.01,0.001)
for(a in alphas) cat( paste("alpha = ",a,"\t|\tpower = ",
           length(power.boot.res[power.boot.res < a])/trials,"\n",sep="" ))
alphas <- seq(0.001,0.5, length.out=1000)
powers <- unlist(lapply(seq_along(alphas),
           function(i)
{
length(power.boot.res[power.boot.res < alphas[i]])/trials
}))
plot(alphas, powers, panel.first=grid(), ylim=c(0,1),
           type="l", col="red", bty="l", main="",
           xlab=expression(alpha), ylab="power")
if(LM) testtype <- c("linear model") else testtype <- c("t-test")
mtext(paste("Simulated power (",testtype,", post-hoc)",sep=""),
           side=3, outer=TRUE, line=0.5, cex=1.4)
mtext(eval(substitute(expression(
           paste("range ",alpha," = [",mialphas,", ",maalphas,
           "] | trials = ",trials,sep="")),
           list(mialphas=min(alphas), maalphas=max(alphas),
           trials=trials))), side=3, outer=TRUE, line=-1, cex=1.1)
ap.tab <- data.frame(alphas, powers)
head(ap.tab)
tail(ap.tab)
```

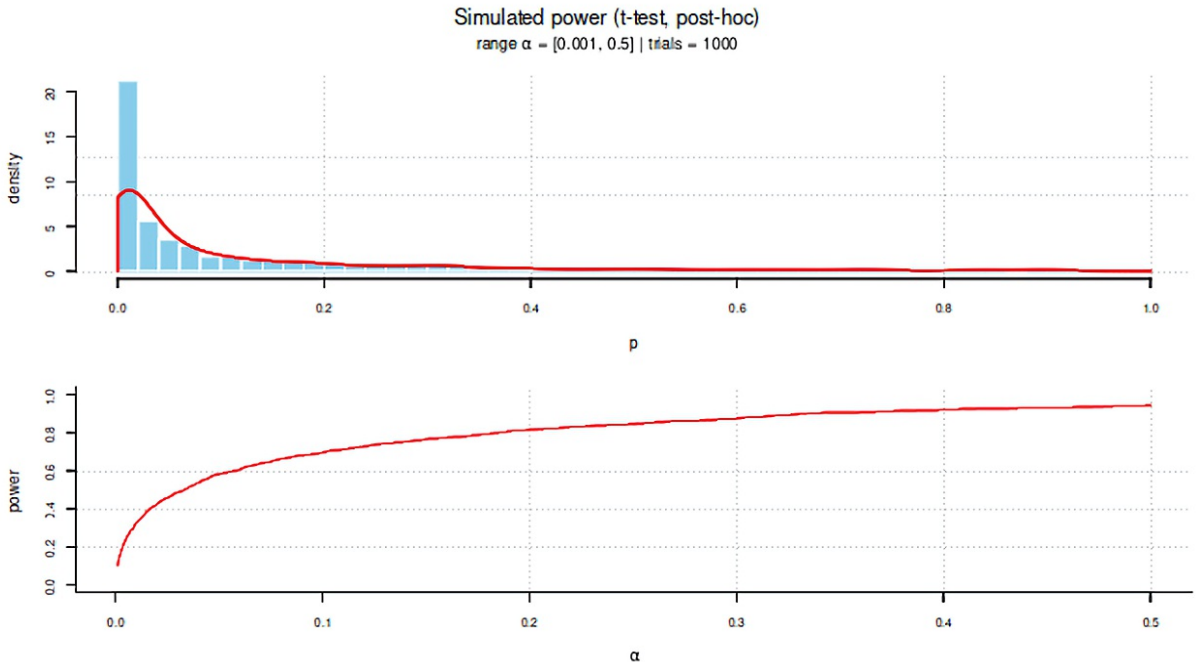


Figura 4.38. *Simulación post-hoc Potencia (prueba t)*

La figura 4.38 correspondiente muestra el histograma y la estimación de la densidad para los *p*-valores simulados en la imagen superior. La imagen inferior muestra la curva de potencia en función del α-nivel. La

curva no es completamente suave, lo que puede atribuirse a la simulación como tal. Como puede verse, la potencia es mayor en los α-niveles más altos cuando el análisis se refiere siempre al mismo conjunto de datos. Esto es fácil de entender, porque a medida que aumenta el nivel α, la barrera para encontrar defectos se hace cada vez más pequeña.y es más fácil encontrarlos. Tomemos como caso individual un α-nivel no convencional predefinido de $\alpha = 0.0678$ como probabilidad crítica de superación, entonces sigue

```
> # only t-test
> length(power.boot.res[power.boot.res < 0.0678])/trials
[1] 0.632
> # Cohen's d
> (6.5-7.8)/2
[1] -0.65
```

y resulta una potencia de 0.632, es decir, en este caso hay una probabilidad de 0.632 de encontrar un efecto del tamaño d = - 0.65 a lo largo de los tamaños especificados anteriormente ($n = 25$, $\sigma = 2$, $\mu_1 = 6.5$, $\mu 2 = 7.8$ así como $\alpha = 0.0678$; $k = 1000$ simulaciones).

> Tarea 4.9: Simulación
>
> La función `sim.DiM()` del ejemplo anterior permite, cuando se pasa LM=TRUE, realizar no una prueba t, sino una regresión entre los dos grupos simulados. La tarea ahora es ejecutar la simulación para LM=TRUE.
> ¿Qué es diferente ahora? ¿Cómo puede explicarse?

Existe un procedimiento comparable en la estadística de Bayes con las Posterior Predictive Checks (PPC – comprobaciones predictivas posteriores; Gelman, Meng & Stern, 1996; Gelman, 2013d), que examina cómo se comportan los datos empíricos en comparación con los datos simulados a partir de la distribución posterior. Aquí el método gráfico demuestra ser fácil de usar y, sin embargo, muy preciso (Gabry, 2018; Gabry, Simpson, Vehtari, Betancourt & Gelman, 2019). Autores como Kruschke (2013c) razonan que los PPC deberían ser completamente bayesianos y que se debería prescindir de los *p*-valores clásicos (véase el capítulo 6.8.4.3).

4.4 Fuentes de error y fenómenos estadísticos

Las estadísticas carecen de una discusión desde la perspectiva de los errores metodológicos típicos de la investigación, los fenómenos estadísticos y las condiciones previas de las pruebas que deben observarse, en las que también pueden producirse sesgos. Sin embargo, no todos los sesgos aparentes son sesgos, como puede verse en el caso de los datos influyentes (véase el capítulo 4.4.12). A menudo no se trata de una cuestión de significancia, sino de hipótesis, suposiciones y tendencias inconscientes por parte de los investigadores para apoyar su propia investigación en lugar de revisarla críticamente. Esto se aplica tanto a la investigación cuantitativa como a la cualitativa. La investigación no mejora con la práctica de una determinada metodología, sino que se aplica según todas las reglas del arte.

En la estadística clásica son posibles muchos sesgos y fuentes de error, además de los fenómenos estadísticos y las condiciones previas, que se exponen temáticamente a continuación. Son los siguientes:

- distorsiones generales (véase el capítulo 4.4.1)
- la búsqueda de significación (véase el cap. 4.4.2)
- el poder (véase el capítulo 4.4.3)
- réplicas (véase el capítulo 4.4.4)
- (auto)engaño (véase el capítulo 4.4.5)
- estimación insesgada de las varianzas muestrales (véase el capítulo 4.4.6)

- aleatorización (véase el capítulo 4.4.7)
- datos que faltan (véase el capítulo 4.4.8)
- procedimientos equivalentes (véase el capítulo 4.4.9)
- hipótesis de distribución normal (véase el capítulo 4.4.10)
- homogeneidad de la varianza (véase el capítulo 4.4.11)
- Valores atípicos y datos influyentes (véase el capítulo 4.4.12)
- potencias de los efectos (véase el capítulo 4.4.13)
- paradojas (véase el capítulo 4.4.14)

4.4.1 Sesgos generales

Los parámetros de un estudio (nivel α- y β, tamaño del efecto, tamaño de la muestra, fuerza de la prueba, dirección de la prueba, dispersión, etc.) son estrechamente interdependientes, de lo contrario no sería posible un análisis de potencia según Neyman-Pearson o un análisis de diseño según Gelman y Carlin (2014). El problema de centrarse únicamente en las significaciones o las medidas asociadas (por ejemplo, intervalos de confianza, análisis de potencia observada, ...), que no utilizan ni generan ninguna información nueva ya se ha tratado (véase el capítulo 4.3.4). Para ilustrar los problemas prácticos asociados, a continuación se presentan algunos escenarios y líneas de pensamiento ficticios y muy simplificados cuando sólo se tienen en cuenta los valores p y se carece de replicación sistemática. Todos los ejemplos apuntan a *fallos de diseño*:

- Una muestra puede ser muy pequeña, de modo que *se pasen por alto efectos relevantes* para la práctica debido a grandes valores p en un único estudio aleatorio extremo. Los resultados no se utilizan de esta forma, por ejemplo, los medicamentos no se desarrollan más.
- A la inversa, pueden encontrarse valores p muy pequeños al azar con una muestra extrema pequeña. En consecuencia, se sobreestima el efecto real, que está presente pero no es relevante en la práctica; por ejemplo, sale al mercado un tratamiento o formación, pero el efecto no está presente o no está claro a quién se aplica y cómo. Las muestras extremas únicas son peligrosas para las generalizaciones. Sin una replicación sistemática y un diseño cuidadoso, las decisiones pueden estar muy viciadas.
- Los datos individuales pueden distorsionar sistemáticamente el panorama general porque se recogieron de forma incorrecta, no contienen información válida, no se recodificaron (por ejemplo, en las pruebas), hubo errores en la introducción de datos, etc. Tales influencias pueden distorsionar masivamente los resultados de los estudios individuales. En este caso, se necesita una buena estrategia para encontrar errores y un profundo conocimiento de cómo se comportan los análisis estadísticos de datos y las variables calculadas en el proceso cuando se producen sesgos. Estos sesgos pueden llevar a realizar pruebas demasiado conservadoras o permisivas, por ejemplo, si las decisiones dependen únicamente de los valores *p*. Los propios valores *p* no nos dicen nada al respecto.
- Ni siquiera los errores de medición conducen siempre a una reducción del tamaño de los efectos. De hecho, lo contrario puede ser cierto para muestras pequeñas y extremas, como informan Loken y Gelman (2017). Así que el lema es que incluso la significación estadística frente a un alto ruido en los datos (= alto error de medición) no significa automáticamente que el efecto encontrado será mayor con un menor error de medición. Es importante examinar adecuadamente la relación señal-ruido. Si el componente de ruido es muy alto y el tamaño de la muestra es pequeño, se recomienda precaución. La significación estadística contiene poca información sobre las razones de su aparición. Todo esto conlleva problemas a la hora de intentar la replicación.
- Los efectos aleatorios siempre pueden estar presentes, independientemente de si se intenta confirmar una hipótesis ("efecto presente") o invalidarla ("efecto ausente"). En ambos casos (falsos positivos y falsos negativos, Ulrich et al., 2016), esto significa que hay pruebas empíricas, pero no pruebas o evidencias en el sentido más estricto del lenguaje cotidiano. Dado que los *p*-valores no son significativos a menos que se tenga en cuenta su contexto de origen y falte una replicación del diseño, las conclusiones quedan prácticamente suspendidas en el aire. El intercambio de hipótesis

nula e hipótesis alternativa conlleva sus propios problemas nuevos, ya que las pruebas de hipótesis no son simétricas.

- La relación entre dos grupos se debe a una tercera variable. Si ésta se controla, la correlación encontrada, que suele ser lineal, desaparece en la nada. La búsqueda de esas terceras variables puede ser difícil, porque los instrumentos de medida pueden incluso no recogerlas. Además, existe el peligro de la paradoja de Simpson (véase el capítulo 4.4.14.1), de modo que no sólo pueden influir las variables de terceros, sino también las características relevantes de subgrupos que interactúan con el objeto.
- No hay que olvidar el pensamiento erróneo. Esto incluye centrarse demasiado en confirmar los propios supuestos que en falsificarlos y cuestionar críticamente los propios supuestos. A menudo se confunden las relaciones correlativas con la causalidad, o se aplican modelos correctos en un contexto equivocado. Los instrumentos pueden ser imprecisos o tener una operacionalización defectuosa. Las estadísticas sólo pueden identificar todos estos factores de forma limitada, si es que lo hacen. Pero todos ellos tienen un impacto más o menos directo en el resultado estadístico y en las conclusiones subsiguientes.

La adecuada consideración de todos estos factores requiere un inmenso esfuerzo, que a menudo sólo puede lograrse de forma limitada, especialmente en el trabajo de cualificación. Obviamente, en este punto se carece de una práctica de investigación a largo plazo y de la experiencia correspondiente. Sin embargo, el problema no radica necesariamente en la falta de experiencia o motivación. Por ejemplo, las propias condiciones del marco empírico pueden limitar un estudio, de modo que se puedan recoger menos datos de los necesarios, se prescinda de repente de variables relevantes o se reduzca innecesariamente la complejidad necesaria para responder a la pregunta de investigación por otros motivos, o simplemente la composición de la muestra cambie de forma desfavorable. Todos estos factores pueden afectar a cualquier estudio, no sólo a los trabajos de cualificación y a la investigación financiada con fondos externos.

Los trabajos de cualificación suelen realizarse en solitario, es decir, por una sola persona, y sólo se dispone de 24 horas al día. Esto y la presión de terminar en un futuro próximo reducen aún más el aprovechamiento del potencial investigador. Incluso los proyectos de doctorado en equipo implican principalmente trabajar en solitario. Los puntos mencionados son sólo ejemplos selectivos de los obstáculos en el camino hacia un estudio perfecto. En última instancia, lo mismo se aplica a los proyectos financiados externamente, sólo que a mayor escala. No todo el mundo puede encontrar tiempo como Lorna Smith Benjamin (1934-) para investigar un único campo de investigación (Structural Analysis of Social Relationships, SASB, Benjamin, 1974) con verdadero detalle durante décadas y elevar así la adquisición de conocimientos a un nivel realmente impresionante. Así, comenzó sus estudios sobre los trastornos de la personalidad ya en 1968, antes de que se incluyeran por primera vez en el DSM-III en 1980 como un trastorno del Eje II. Benjamin publicó por primera vez sus ideas sobre el SASB como artículo en 1974, y su obra fundacional sobre el diagnóstico interpersonal y el tratamiento de los trastornos de la personalidad apareció en 1993. Y, obviamente, ya llevaba mucho tiempo trabajando en ello. La tendencia actual hacia la unidad publicable más pequeña, con el fin de publicar tanto como sea posible en poco tiempo y en diferentes revistas, hace que la ciencia no solo sea ineficiente y poco económica (navaja de Occam, Baker 2016), sino además poco fiable y poco científica. Además, dificulta innecesariamente que otros científicos se familiaricen con un tema.

Del mismo modo que no se da por sentado que las teorías y los modelos sean correctos en absoluto – "todos los modelos son erróneos", aforismo popular en filosofía de la ciencia y estadística – nadie cree seriamente que exista un estudio de investigación perfecto. La autorreflexión crítica y el intercambio con otros ocupan un lugar importante en el proceso científico como imperativo. No informar a sabiendas (por ejemplo, porque no son significativas) es tan erróneo como informar mal a sabiendas o incluso inventarse cosas. Solo de los errores se pueden extraer lecciones. Una cultura del error creativo sin consecuencias para las posibilidades de publicación sigue siendo el problema central de las próximas generaciones de investigadores. Precisamente porque esto simplemente no existe todavía en la práctica de publicación relevante. En la actualidad, lo que se publica es lo que "algo" descubre, y "algo" se define de tal manera que en última instancia debe haber una significación estadística detrás. Con respecto a los modelos, sin embargo, es importante comprender qué aspectos de los datos pueden describir adecuadamente y cuáles no, y no probarlos o su contrario (Gelman, 2007a; 2008b).

Tarea 4.10: Publicaciones y política

El lector interesado podría plantearse desarrollar algunos escenarios propios o reflexionar a partir de su propia práctica investigadora sobre cómo se manejan los valores p o las significaciones y qué es lo que específicamente puede o incluso ha llevado a su sesgo. Además, deberían revisarse las propias investigaciones anteriores y cuestionarse todas las decisiones en el marco de estos procesos de investigación para saber por qué salieron como salieron y no de otra manera. ¿Qué podría haber sido diferente si tal o cual decisión se hubiera tomado de otro modo? ¿Qué habría sido mejor, qué posiblemente peor? ¿Habría habido una solución perfecta? Y, sobre todo, ¿qué hacer con estas conclusiones para el próximo proyecto de investigación?

Esto nos lleva a la siguiente tarea: reflexionar sobre el proceso de presentación de publicaciones o resúmenes a congresos, por qué se aceptaron o no y qué cambios hubo que hacer para que se aceptara un artículo o un resumen.

4.4.2 En busca de significados - intenciones de investigación inconscientes y *p*-hacking

La significación como tal es sólo un componente del razonamiento estadístico y no el factor más importante. Hoy en día son posibles muchas variantes para trabajar estadística seria y clásicamente al margen de la significación, si se tienen en cuenta los problemas epistemológicos de la práctica actual que se han mencionado. Por ejemplo, en el caso de los modelos de regresión, Gelman y Hill (2007) sugieren no utilizar los valores *p* para decidir si los parámetros permanecen en el modelo. En su lugar, recomiendan dejar los predictores en el modelo si apuntan en la dirección correcta (error de tipo S), tienen el tamaño adecuado (error de tipo M) y son teóricamente sólidos (véase el análisis del diseño en el capítulo 4.3.3.2). En la práctica, los autores recomiendan dejar los coeficientes de regresión en el modelo si sus valores *t* son aproximadamente ≥ 2, lo que corresponde aproximadamente a un nivel del 5 % en el sentido clásico. Sin embargo, esto sólo se aplica si la dirección y la magnitud de los efectos son coherentes con las consideraciones teóricas. En consecuencia, `display()` en el paquete R `arm` de los autores no muestra valores *t* o incluso *p*, sino sólo coeficientes y errores estándar. Los valores *t* pueden generarse manualmente dividiendo coeficientes/errores estándar. Los valores *p* necesarios pueden calcularse fácilmente a partir de los coeficientes y sus errores estándar, si es que son necesarios. Por otro lado, los autores recomiendan utilizar su función `sim()` para simular datos que se ajusten al modelo y obtener errores estándar para las estimaciones de los parámetros (`ptII_quan_classicstats_p-hacking-sim.r`).

```
> # without and with p-values
> display(lm.fit)
lm(formula = weight ~ group)
coef.est coef.se
(Intercept) 5.03 0.22
groupTrt -0.37 0.31
---
n = 20, k = 2
residual sd = 0.70, R-Squared = 0.07
> display(lm.fit, detail=TRUE)
lm(formula = weight ~ group)
coef.est coef.se t value Pr(>|t|)
(Intercept) 5.03 0.22 22.85 0.00
groupTrt -0.37 0.31 -1.19 0.25
---
n = 20, k = 2
residual sd = 0.70, R-Squared = 0.07
> lm.fit.sim <- sim(lm.fit)
> str(lm.fit.sim)
Formal class 'sim' [package "arm"] with 2 slots
```

```
..@ coef : num [1:100, 1:2] 4.98 4.59 5.21 5.16 5.33 ...
.. ..- attr(*, "dimnames")=List of 2
.. .. ..$ : NULL
.. .. ..$ : chr [1:2] "(Intercept)" "groupTrt"
..@ sigma: num [1:100] 0.686 0.784 0.748 0.762 0.568 ...
```

`sim()` se basa en bootstrap paramétrico (Efron, 1979) extrayendo los β-pesos del modelo lineal de una distribución normal multivariante y generando los errores estándar de los β-pesos utilizando valores aleatorios de la distribución χ^2 (véase el código R de `sim()` en `arm`), cada uno basado en los parámetros estimados del modelo empírico. El código R de las funciones dentro de los paquetes R puede obtenerse con un esfuerzo razonable:

```
# see R source code R-Code
arm:::sim
showMethods("sim")
?getMethod
getMethod("sim", "lm")
```

Los análisis más complejos, como los de `lmer()` o `glmer()` en el paquete R `lme4` para analizar HLMs/MLMs, no muestran valores *p* en absoluto. Según el autor del paquete, Bates (2006), no está claro exactamente cómo se calculan los grados de libertad, por lo que los valores *p* resultantes estarían sujetos a una incertidumbre muy grande (Robinson, 2008; Bolker et al., 2019). En consecuencia, no se implementan y deben ajustarse mediante módulos externos post-hoc con la incertidumbre correspondiente antes mencionada. Entonces sigue sin estar claro qué es exactamente lo que pueden decir.

Sin embargo, este no es ni mucho menos el final de la discusión sobre los *p-valores extremos*. Así, Gelman et al. (2004, p.175f.), por ejemplo:

> „A model is suspect if a discrepancy is of practical importance and its observed value has a tail-area probability that is close to 0 or 1, thereby indicating that the observed pattern would be unlikely to be seen in replications of the data if the model were true. An extreme p-value implies that the model cannot be expected to capture this aspect of the data. ... In some cases, even extreme p-values may be ignored if the mis t of the model is substantively small compared to variation within the model. ... If a p-value is close to 0 or 1, it is not so important exactly how extreme it is. A p-value of 0.00001 is virtually no stronger, in practice, than 0.001; in either case, the aspect of the data measured by the test quantity is inconsistent with the model. A slight improvement in the model (or correction of a data coding error!) could bring either p-value to a reasonable range (between 0.05 and 0.95, say). The p-value measures 'statistical significance,' not 'practical significance.'
>
> The relevant goal is not to answer the question, 'Do the data come from the assumed model?' (to which the answer is almost always no), but to quantify the discrepancies between data and model, and assess whether they could have arisen by chance, under the model's own assumptions. [...] Finding an extreme p-value and thus 'rejecting' a model is never the end of an analysis [...]"

Aprendemos: los p-valores extremos no son de fiar en ningún campo. En las ciencias sociales hay muchos escenarios (por ejemplo, experimentos, hipótesis de modelos y los correspondientes modelos estadísticos, etc.) en los que paradójicamente la hipótesis es que la hipótesis nula es cierta. Esto significa que el objetivo de la prueba es entonces precisamente no rechazar la hipótesis nula mediante los datos empíricos. Sin embargo, no rechazar la hipótesis nula no implica su validez, sólo que el modelo alternativo no es superior. El mantenimiento de la hipótesis nula no es una prueba de su validez– una situación difícil desde el punto de vista de la investigación. Este tipo de situaciones pueden darse en diseños experimentales (por ejemplo, pruebas de equivalencia) o en la aplicación de modelos de ecuaciones estructurales. A partir de esta situación Rost (2003, párrafo 23) ve un grave problema de la aplicación estadística actual en el

> "Mientras que en la estadística inferencial clásica, según NEYMAN y PEARSON, la hipótesis nula suele especificar el modelo que, según la teoría científica, no se espera y que, por tanto, se desea

> rechazar (por ejemplo, la hipótesis nula), la hipótesis alternativa suele especificar el modelo que, según la teoría científica, no se espera y que, por tanto, se desea rechazar (por ejemplo, la hipótesis alternativa). (por ejemplo, la igualdad de los valores medios o la ausencia de correlación lineal), en la inferencia "moderna correlación), en las comprobaciones de validez de modelos "modernas" el modelo de interés está en la hipótesis nula. está en la hipótesis nula. Esta inversión tiene implicaciones para la especificación y el control del error alfa y beta...".

Este problema ya se ha puesto en práctica en los experimentos de reflexión del capítulo 4.3.3.4 sobre los dos estudios de casos de la gestión de la calidad en el contexto de la teoría de Neyman-Pearson, en los que se eligieron niveles de error muy diferentes para α y β en cada caso. Se demostró que la asignación de las condiciones a las respectivas hipótesis puede dar lugar a cambios drásticos. Por tanto, no se trata sólo de valores *p* extremos como propiedades de los datos, sino de la contribución del contenido de las hipótesis H_1 o H_2 y en qué dirección deben, por tanto, ponerse a prueba.

Los problemas reales en la práctica diaria de la investigación rara vez giran en torno al puro *p-hacking* o *p-shing* (Ulrich et al., 2016) *p*-valores en los análisis de tal forma que se cree significancia. El ni siquiera tiene que ser intencionado, como mostraremos a continuación. Muy pocos investigadores manipulen deliberadamente los datos. Más bien hay que suponer una cierta ingenuidad acerca de cómo se relaciona una recogida de datos poco seria con hallazgos supuestamente significativos.

Entonces es importante exigir estudios de replicación realistas para reducir la dependencia de la muestra de los datos en los estudios individuales. Pero incluso esto no es necesariamente suficiente, como muestra Bicare (2016) utilizando una aplicación web. Resulta aún más difícil cuando la investigación está estructurada de tal manera que habitualmente solo conduce a la confirmación de los propios supuestos en lugar de cuestionarlos activa y críticamente en el sentido de Popper. Exactamente tales supuestos residen lejos de la argumentación puramente estadística, pero tienen un impacto directo en ella. Dos casos prácticos de estudios empíricos ayudan a ilustrar el problema: el estudio "50 Shades of Grey" (Nosek, Spies y Motyl, 2012) y el estudio de Bem (2011a) sobre la clarividencia.

4.4.2.1 El estudio "Fifty-Shades-of-Gray"

Nosek, Spies y Motyl (2012) realizaron un estudio sobre la percepción y el juicio en el contexto de las actitudes políticas: el estudio "50 sombras de gris". Encontraron una relación estadística altamente significativa entre el extremismo político y la percepción de imágenes (blanco/negro frente a tonos de gris). En lugar de publicar este resultado conforme a la teoría, replicaron su propio estudio. Basándose en el primer estudio, tenían un 99% de posibilidades de volver a obtener un valor *p* de $p < 0{,}05$. Sin embargo, la replicación arrojó un valor *p* de $p = 0{,}59$. Lo más difícil en este caso no es publicar inmediatamente, sino esperar a la replicación. ¡Respeto a los autores!

4.4.2.2 El estudio de Bem sobre la clarividencia

El psicólogo social estadounidense Daryl Bem (2011a, véase también el análisis del diseño en el capítulo 6.8.1.6) intentó aportar pruebas de la percepción extrasensorial en estudiantes universitarios, comúnmente conocida como "clarividencia" o ESP = extra-sensory-perception, Psi o también Clairvoyance. Entramos en más detalles teóricos y metodológicos en el capítulo 6.8.1.6 en forma de un excursus sobre el diseño. Este estudio se utiliza a menudo de forma crítica, no sólo en el contexto de la estadística de Bayes, y es criticado por muchos autores como un ejemplo de ciencia mala y no replicada. En su experimento, Bem encontró resultados estadísticamente significativos para las imágenes eróticas frente a las no eróticas, que Bem intentó explicar en términos evolutivos de ventaja reproductiva. Así, se supone que la predicción de una pareja potencial aporta ventajas en la reproducción sexual – en resumen, traer más hijos al mundo y propagar más los propios genes. Sin embargo, la ventaja reproductiva es un argumento asesino en el campo evolutivo y no puede estudiarse directamente porque los periodos de tiempo son demasiado largos. Visto a posteriori, la ganancia de conocimiento es manejablemente pequeña: los que todavía están por aquí aparentemente tuvieron una ventaja evolutiva y más descendencia. Esto hace que la explicación subyacente sea muy débil,

ya que no se puede falsar prácticamente; y apenas cumple los requisitos de una teoría científicamente comprobable.

No obstante, el estudio fue publicado por el *Journal of Personality and Social Psychology (JPSP)*, una revista muy respetada hasta la fecha (Bem, 2011a). El estudio, sin embargo, no pudo ser replicado por otros investigadores independientes y también recibió considerables críticas (Wagenmakers, Wetzels, Borsboom & van der Maas, entre otros, 2011; Rouder & Morey, 2011; Rouder, Morey & Province, 2013) y desencadenó un amplio debate metodológico que continuó en blogs (Yarkoni, 2011; Schimmack, 2015b; Gelman, 2019c) y en artículos de revistas (Ritchie,Wiseman & French, 2012; Francis, 2012; Traxler, Foss, Ruchira & Zirnstein, 2012). Los argumentos dieron lugar a contraargumentos (Bem, Utts & Johnson, 2011) y así sucesivamente.

Sin embargo, más significativas parecen ser las dificultades experimentadas por varios investigadores para publicar sus elaboradas réplicas (French, 2012). Entre ellos, Snodgrass (2011-09-27), Galak (2010-10-09) y Galak, LeBoeuf, Nelson & Simmons (2012-06-19), respectivamente. Aún más interesante resulta la noticia aparecida en el periódico *DER SPIEGEL ONLINE* (Weber, 2012-03-15) de que Bem, aunque clasificó el estudio como fácilmente replicable, luego bloqueó activamente uno de los intentos de replicación en el *British Journal of Psychology* en calidad de revisor, alegando que los investigadores no creían en las habilidades paranormales y, por tanto, no estaban cualificados para hacerlo: "Alguien más debería dirigir la replicación".

La *Journal of Personality and Social Psychology* rechazó las réplicas incluso antes de la revisión, declarando que nunca publicarían réplicas (Aldhous, 2011-05-05). Por lo tanto, la revista se ha expuesto como defectuosa y Bem no sale bien parado en absoluto, independientemente de los defectos científicos.

Al margen de esta escaramuza política, el estadístico estadounidense Andrew Gelman y su colega Eric Loken resumen los subyacentes con su colega Eric Loken, resume el problema subyacente sobre todo desde un punto de vista metodológico de la investigación y no solo desde una perspectiva estadística (Gelman & Loken, 2013, 2014):

> „Bem's paper presented nine different experiments and many statistically significant results — multiple degrees of freedom that allowed him to keep looking until he could find what he was searching for. But consider all the other comparisons he could have drawn: If the subjects had identified all images at a rate statistically significantly higher than chance, that certainly would have been reported as evidence of ESP. Or what if performance had been higher for the nonerotic pictures? One could easily argue that the erotic images were distracting and only the nonerotic images were a good test of the phenomenon. If participants had performed statistically significantly better in the second half of the trial than in the first half, that would be evidence of learning; if better in the first half, evidence of fatigue."

Por este motivo, las teorías y las hipótesis, así como las condiciones marco de los estudios (por ejemplo, el tamaño y la composición de la muestra), deberían formularse con precisión de antemano y no a posteriori. (por ejemplo, el tamaño y la composición de la muestra) deberían formularse idealmente con precisión de antemano y no después y no a posteriori, es decir, cuando aún no se dispone de datos. Asimismo, el proceso también debe documentarse meticulosamente. No sólo en la estadística clásica no sólo en la estadística clásica la recogida de datos debe planificarse meticulosamente con antelación para que pueda llevarse a cabo exactamente de la misma manera. Después siempre hay razones por las que algo "tenía" que hacerse de esta manera y no de otra. A diferencia de la estadística bayesiana, en la estadística clásica un conjunto de datos existente no puede utilizarse directamente en el análisis del siguiente conjunto de datos. Los datos son productos aleatorios y pueden dar lugar a metaanálisis o formar el punto de partida de un análisis de potencia para futuros escenarios de investigación. Una integración directa de conjuntos de datos no está prevista. Siempre existe el peligro de una α-acumulación y deben tomarse contramedidas en consecuencia. En la estadística este problema no existe. En el análisis de nuevos datos, los datos existentes pueden, en principio, ser información a priori directamente en el teorema de Bayes y las ecuaciones y contribuir así a un resultado más preciso. Esto corresponde al modelo de "aprendizaje a partir de la experiencia acumulada", que idealmente se ejecuta hasta que las ecuaciones se estabilizan y no se obtienen más conocimientos a través de nuevos datos. No existe un problema equivalente a la α -acumulación. Sin embargo, no se puede concluir de ello que la práctica de la estadística bayesiana vaya siempre acompañada de réplicas y del procedimiento

descrito de *aprendizaje a partir de la experiencia*, o que per se la aplicación metodológica aquí sería fundamentalmente mejor o que la explicación del conocimiento previo se hace tan cuidadosamente que realmente toda la información contextual disponible se recoge, reconstruye, pondera y sólo entonces da lugar a una o más distribuciones numéricas. Cabe sospechar que éste no es precisamente el caso. Ya existen diversas tendencias a construir barreras fijas y rígidas en el campo bayesiano, que pueden ser tan obstructivas como los valores *p* y las significaciones, y hay investigadores que advierten de ello (Gigerenzer & Marewski, 2015). Discutiremos el estudio de Bem más adelante, en el capítulo 6.8.1.6, desde un punto de vista metodológico y justificaremos que, aunque *todas* sus hipótesis sobre fenómenos paranormales existieran, hizo un trabajo (teórico) muy pobre, y en consecuencia, por el diseño elegido y la selección de la muestra, no puede salir nada significativo de él.

Autores como Bicare (2016) o sitios web como FiveThirtyEight (2019) ofrecen pequeñas apps web con las que se puede simular el p-hacking de forma muy performante. Para una primera impresión, el siguiente script de R es suficiente para demostrar el problema básico (`ptII_quan_classicstats_p-hacking-sim.r`). El script R simula dos distribuciones y aumenta sucesivamente el tamaño de la muestra hasta que se obtiene un resultado estadísticamente significativo mediante una prueba t simple. Ignoramos en el script la necesaria corrección del umbral crítico de rebasamiento debido a la repetición de pruebas, ya que estamos hablando de p-hacking y simplemente seguimos probando hasta que nos salte un resultado estadísticamente significativo. Puedes repetir esto muchas veces y hacerte una idea de lo grande que tiene que ser normalmente una muestra para obtener un efecto de un cierto tamaño de efecto a un cierto nivel de tasa de error. Esto muestra cómo se pueden obtener resultados estadísticamente significativos aumentando sin sentido el tamaño de las muestras o por la forma en que se establece el corte del tamaño de una muestra. Y como recordatorio - aquí estamos tratando datos puramente aleatorios.

```
> seed <- 0798
> set.seed(seed)
> # further explanations:
> # N|group = N per/ each group
> # N|addon = additional N
> # s0 = sigma0
> # pv = p-value
> # d|true = true d (theory)
> # d|pool.sd = pooled sd (d)
> res1 <- p.hack.sim(pr=TRUE)
     no N|group N|addon mu0 mu0+d s0   alpha pv   d|true  d|pool.sd
[1,] 1  30      0       0   0.1   1    0.05  0.79 0.1    -0.070
[2,] 2  35      5       0   0.1   1    0.05  0.99 0.1    -0.003
[3,] 3  40      5       0   0.1   1    0.05  0.98 0.1    -0.006
[4,] 4  45      5       0   0.1   1    0.05  0.82 0.1     0.047
[5,] 5  50      5       0   0.1   1    0.05  0.83 0.1     0.043
[6,] 6  55      5       0   0.1   1    0.05  0.44 0.1     0.147
[7,] 7  60      5       0   0.1   1    0.05  0.38 0.1     0.161
[8,] 8  65      5       0   0.1   1    0.05  0.20 0.1     0.227
[9,] 9  70      5       0   0.1   1    0.05  0.11 0.1     0.270
[10,]10 75      5       0   0.1   1    0.05  0.11 0.1     0.261
[11,]11 80      5       0   0.1   1    0.05  0.04 0.1     0.323
Two-sample t test power calculation
              n = 1571
          delta = 0.1
             sd = 1
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
NOTE: n is number in *each* group
```

Es significativo que el tamaño de la muestra calculado a priori mediante `power.t.test()` se redujera considerablemente en el ejemplo anterior para lograr un resultado estadísticamente significativo. En lugar de una muestra de N = 1571, sólo fue necesario un tamaño de muestra de N = 80, es decir, una reducción por un factor de 19:6 para encontrar el efecto deseado de d = 0.1 al nivel habitual del 5%. En este ejemplo no tuvimos que probarlo muy a menudo. Esto ayuda a comprender qué es una muestra aleatoria y qué consecuencias tiene para los estudios *no* repetidos. Además, el ejemplo ilustra que una potencia objetivo de,

digamos, 0.8 significa identificar típicamente un efecto según la potencia elegida con los cálculos a priori asociados (tamaño de la muestra, tamaño del efecto, niveles de tasa de error, etc.). Esto puede ocurrir – aleatoriamente o no – con muestras aún más pequeñas o no con muestras más grandes. La certeza real es limitada, incluso con mediciones repetidas. Así lo demuestra, por ejemplo, la discusión sobre el estudio de Bem. Sin embargo, se puede deducir de las indicaciones empíricas cierta plausibilidad como se debe evaluar la seriedad de un resultado estadístco.

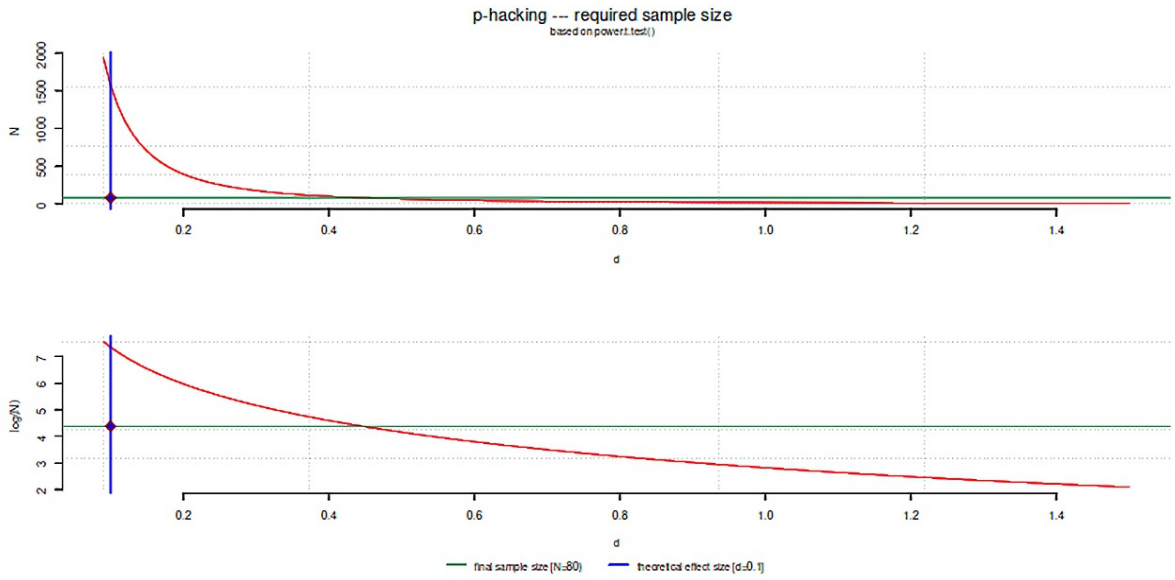


Figura 4.39. *Simulación* p*-hacking (Tamaño de muestra)*

Podemos repetir estas simulaciones varias veces para hacernos una idea aproximada del tamaño que debe tener normalmente una muestra para alcanzar la significación convencional en la práctica (simulada)... Ejecutamos 1000 réplicas con lo siguiente llamada a `p.hack.sim()`:

```
sim.res <- p.hack.sim(d=0.2, seed=seeds[x], sigma0=1, alpha=0.05,
pr=FALSE, graph=FALSE, ppaur=FALSE)
```

Podemos compararlo con un análisis de potencia a priori mediante `power.t.test()`.

```
> ptt <- power.t.test(n=NULL, delta=0.2, sd=1, Output
+ sig.level=0.05, power=0.8,
+ type="two.sample",
+ alternative="two.sided", strict=TRUE)
> #n = number of observations (per group)
> ptt
Two-sample t test power calculation
n = 393
delta = 0.2
sd = 1
sig.level = 0.05
power = 0.8
alternative = two.sided
NOTE: n is number in *each* group
> describes(sim.mat)
N (all)     NA   N (no NA) sum     mean g  mean h
no steps    1000 0 1000      30089  30.1   Inf
samplesize 1000 0 1000     175445 175.4   Inf
 h           mean    var    sd vc  med    mad  min    max
no steps     4.82   1066  32.7  1.085  19   23.7  1 292
samplesize 79.28  26654 163.3  0.931 120  118.6 30 1485
```

```
range      mean  dev   skewness kurtosis 1.Quantil
no steps    291   24.5 1.95     6.34      5
samplesize 1455  122.7 1.95     6.34     50
3.Quantil  IQR SE  mean CI low mean CI up mean
no steps    44  39 1.03  28.1        32.1
samplesize 245 195 5.16 165.3       185.6
> summary(sim.mat)
        no steps samplesize
Min. :    1.0     Min. :  30
1st Qu.:  5.0   1st Qu.:  50
Median : 19.0   Median : 120
Mean :   30.1     Mean : 175
3rd Qu.: 44.0   3rd Qu.: 245
Max. :  292.0     Max. :1485
> fivenum.wn(sim.mat)
Min 1st Qu. Median 3rd Qu. Max
1   19      46     130     1485
```

lo que sigue al resultado siguiente (véase figura 4.40):

```
> # compare with theoretical value by power.t.test()
> # percent of trials required that are less
> # than the theoretical value
> # (= required N / sample size) but nevertheless
> # attain 'statistical significance'
> max(sim.mat.ssize.cs.pc[as.numeric(
+ names(sim.mat.ssize.cs.pc)) <= ptt$n])
[1] 90.2
```

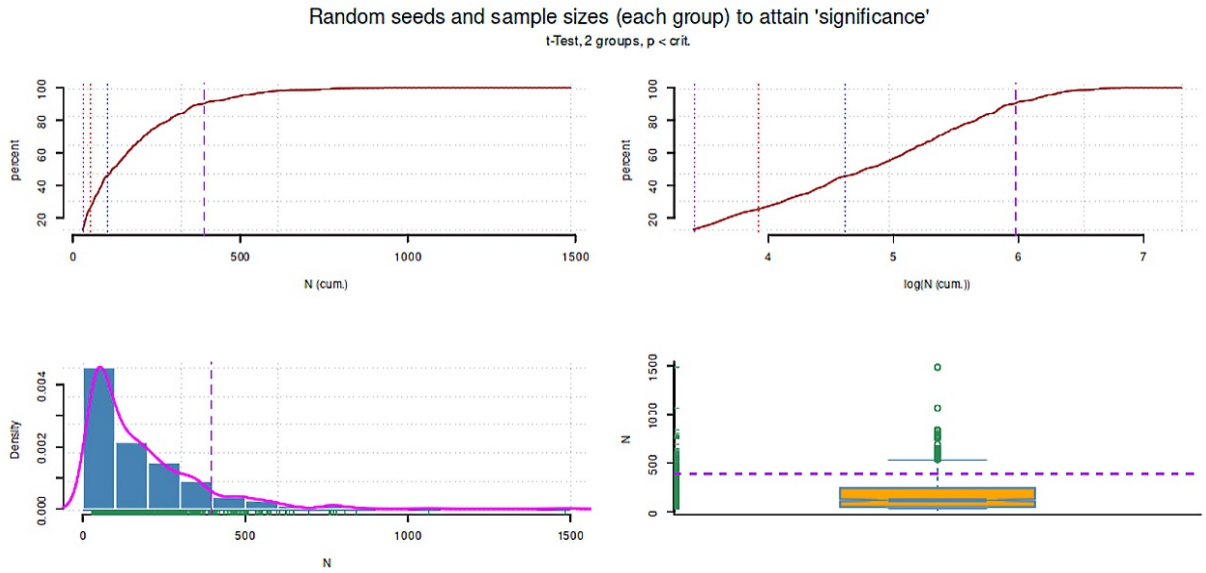


Figura 4.40. *Simulación p-hacking (aleatoriedad, tamaño de la muestra y significación)*

Para decirlo sin rodeos, se pueden derivar algunas pautas para lograr un resultado significativo con un alto grado de certeza (véase también Bicare, 2016):

- Es mejor realizar muchos estudios con pocos participantes que un estudio con muchas personas.
- Mejor elegir un tamaño del efecto lo más pequeño posible, tanto mejor si aquí se encuentra algo, ya que se puede demostrar crear algo cuasi de la nada.
- Se recomienda el uso de muchas variables dependientes.

- La exclusión de valores atípicos mediante el análisis `boxplot` y la eliminación de otros puntos de datos que interfieren tiene un efecto positivo – siempre se pueden encontrar razones para ello.
- A continuación, hay que concentrarse en las variables que están cerca del nivel de significación y todas las demás pueden ignorarse por el momento.
- Si esto no es suficiente, se añaden participantes al estudio paso a paso, controlando la edad y el sexo, por ejemplo.
- Ahora se comprueba una y otra vez si no se alcanza el umbral de significación. Si es así, se puede guardar el estudio y empezar de nuevo. El mismo esquema se aplica, ya que las revistas de hoy en día esperan réplicas o estudios múltiples.

Por su parte, Simonsohn, Nelson y Simmons (2014, 2014-01-11) y Simmons y Simonsohn (2017), por ejemplo, ofrecen la aplicación web P-curve (Simonsohn, Nelson & Simmons, 2013) como herramienta para examinar la seriedad de los *p*-valores. Las hipótesis deben formularse de forma restrictiva y sin ambigüedades. De lo contrario, los resultados encajarán de algún modo "casual" en el edificio de teorías más o menos conscientemente favorecido por los investigadores (palabra clave: tasa de falsos descubrimientos, véase el Cuadro 4.2, p.78). Asimismo, ya existe una teoría para casi todo, por lo que el complejo proceso de traducción de las afirmaciones relacionadas con el contenido a las hipótesis estadísticas y viceversa (Gigerenzer, 1981; ampliación para datos cualitativos en Gürtler, 2005, véase también el capítulo 7.11) debe examinarse con especial cuidado. No hay que olvidar que las correlaciones nunca son relaciones causa-efecto y las terceras o cuartas variables influyentes pueden dar lugar a sesgos si no se controlan metódicamente. A veces faltan variables de agrupación significativas, por lo que los efectos pueden aparecer o no debido a la composición de la muestra, aunque los mismos efectos parezcan completamente diferentes en los subgrupos pertinentes. Esto significa que las decisiones más o menos conscientes tomadas durante un proceso de investigación pueden conducir a resultados que se ajusten a las hipótesis, sin que los investigadores puedan ser acusados de una estrategia intencional dirigida a *p*-shinging o *p*-hacking (Gelman y Loken, 2014). En realidad, esto lo empeora, ya que estos procesos tienden a ser implícitos e inconscientes, como señalan Gelman y Loken (2013, p.10), y por eso la línea entre causas y efectos, hipótesis y conclusiones se difumina:

> „In this garden of forking paths, whatever route you take seems predetermined, but that's because the choices are done implicitly. The researchers are not trying multiple tests to see which has the best p-value; rather, they are using their scientific common sense to formulate their hypotheses in a reasonable way, given the data they have. The mistake is in thinking that, if the particular path that was chosen yields statistical significance, this is strong evidence in favor of the hypothesis."

4.4.3 La gestión del poder

La potencia o poder discriminatorio (también denominado fuerza, bondad o potencia de la prueba) es una propiedad de la prueba en la teoría de las pruebas y no una propiedad de los datos como el valor *p*. La potencia se define como la capacidad de una prueba para rechazar correctamente la H_0 a favor de una alternativa específica y verdadera H_1, es decir, para tomar la decisión correcta y reconocer o identificar correctamente un efecto como tal. Los valores de la potencia oscilan entre cero y uno. Al aumentar la potencia, disminuye la probabilidad de un error de tipo II (= suposición de una hipótesis nula falsa, véase la Tabla 4.2, p.78) o la tasa de error β (= retención o suposición errónea de H_0 , ya que: Potencia = 1 - β).

Según las convenciones imperantes (Cohen, 1969, 1992), se elige un valor aproximadamente cuatro veces mayor que el nivel de significación. Por ejemplo, un nivel de convención de $\alpha = 0.05$ da como resultado $\beta = 0.20$ y, en consecuencia, una potencia = 0.80. Tanto la elección de α como la relación de α a β no deben hacerse de acuerdo con las convenciones, como se subraya a lo largo de este libro. A diferencia de α no se puede dominar directamente β – y si, entonces mediante un análisis de potencia priore con planificación de muestras. La potencia depende de varios factores. Entre ellos

- Prueba y diseño – cuanto más exacta, precisa y probada sea una prueba, mayor será la potencia para encontrar de forma fiable cualquier efecto. Lo mismo se aplica al diseño. Cuanto más preciso sea el diseño, mayor será la probabilidad de encontrar efectos existentes.
- Tamaño del efecto δ – cuanto mayor sea el tamaño del efecto, más rápida y fiablemente se encontrará un efecto.
- Error de muestreo σ – cuanto menor sea, más precisa será la medición y más seguro será encontrar efectos existentes.
- Tasa de error α – cuanto mayor, mayor será la potencia, a menos que no se haya fijado β de antemano.
- Tasa de error β – cuanto menor sea, mayor será la potencia, siempre que se haya fijado β y se lleve a cabo la correspondiente planificación previa de la potencia en el sentido de Neyman-Pearson para determinar el tamaño óptimo de la muestra.
- Tamaño de la muestra N – cuanto mayor sea, menor será el error de muestreo y más precisa será la medición, es decir, aumenta la potencia.
- Dirección de la hipótesis – las pruebas bilaterales requieren una muestra mayor que las unilaterales o tienen una potencia menor con el mismo tamaño de muestra. Por otro lado, las pruebas unilaterales y direccionales requieren una buena fundamentación del contenido.
- Nivel de escala – una prueba paramétrica tiene mayor potencia que una prueba no paramétrica para el mismo tamaño de muestra N, pero sólo si las distribuciones supuesta y verdadera coinciden. Si las distribuciones supuestas no coinciden con la realidad, una prueba no paramétrica puede tener mayor potencia que una paramétrica.
- Gravedad – según John, Loewenstein y Prelec (2012), la gravedad puede dividirse en intencionada frente a ignorancia. En el primer caso, ningún consejo metodológico será de ayuda. En el segundo caso, la potencia observada debería acercarse lentamente a la potencia real en las réplicas. La investigación dudosa sobreestima la potencia debido a técnicas diferentes y cuestionables. Estas incluyen el énfasis excesivo en resultados significativos en muestras pequeñas y altos niveles de ruido en los datos, p-hacking (Simmons, Nelson & Simonsohn, 2011; Head, Holman, Lanfear, Kahn & Jennions, 2015), reglas de detención poco claras en la recopilación de datos, etc., todo lo cual distorsiona artificialmente la potencia y los efectos reales encontrados en la población.

Dado que la potencia real o verdadera no puede medirse ni controlarse directamente (véase la teoría de la prueba clásica), debe inferirse. Puede determinarse mediante metaanálisis de la mediana de la potencia observada (Brunner & Schimmack, 2018a) o, en parte, mediante estudios de simulación. La potencia post-hoc observada de un único estudio suele estar sesgada, aunque únicamente por artefactos de muestreo. Schimmack (2015a) asume, sin embargo, que a medida que aumenta el número de estudios disponibles, la mediana de la potencia observada se aproxima bien a la potencia verdadera, ya que la mediana siempre tiene una probabilidad de 50:50 de sobreestimar o subestimar la potencia verdadera (véase sobre la determinación de la potencia post-hoc, Yuan & Maxwell, 2005 o Schimmack, 2015a). En este caso, el valor medio no sería verdadero desde el punto de vista de las expectativas y sobrestimaría sistemáticamente la potencia real en valores altos y subestimaría sistemáticamente la potencia real en valores bajos. En los estudios de simulación o en el caso de varios estudios existentes, la potencia se calcula mediante el número de resultados estadísticamente significativos qua recuento. Este método también es adecuado para los metaanálisis. El número de resultados estadísticamente significativos se suma y se divide por el número de estudios publicados para determinar la potencia existente. De este modo, se determina la potencia bien mediante

- una estimación matemática en función del tamaño de la muestra, el tamaño del efecto, etc. a priori según Neyman-Pearson, o bien
- utilizando la potencia de los resultados estadísticamente significativos de los estudios existentes en relación con el número total de estudios realizados.

Ambos valores son idénticos si se publican todos los resultados y no solo selectivamente y no se produce p-hacking (Bittner-Stephan, 2015). Brunner y Schimmack (2018a, 2018b) derivan una variable de prueba (*RIndex*, Schimmack, 2016a) a partir de la discrepancia entre la potencia comunicada y la real con el fin de evaluar la replicabilidad y, por tanto, la seriedad de los resultados de los estudios (véase también la discusión

en Hughes, 2015-01-29). La práctica de publicación sesgada de las revistas científicas, que favorecía las significaciones y esta práctica no ha cesado en absoluto, tiene un efecto especialmente problemático. Así, para los meta-estudios, en realidad debería añadirse automáticamente un cierto porcentaje de estudios insignificantes y no publicados, para introducir un factor de corrección razonable. En un estudio diseñado según Neyman-Pearson, la potencia objetivo resulta del diseño, el tipo de análisis de los datos (tipo de prueba, como la prueba *t* y, en consecuencia, si es unilateral o bilateral, o si se trata de una, dos muestras o dos muestras pareadas, etc.)., el nivel y el tamaño del efecto objetivo. Ahora un ejemplo de R.

```
> power.t.test(n=30, delta=1.1, sd=2.9, sig.level=0.01, power=NULL,
type="paired", alternative="two.sided")
Paired t test power calculation
n = 30
delta = 1.1
sd = 2.9
sig.level = 0.01
power = 0.269
alternative = two.sided
NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs
```

Los resultados de la potencia indican que se basan en una hipótesis bilateral, una muestra emparejada (n indica aquí el número de pares), una diferencia media real de $\delta = 1.1$, una desviación estándar (aquí: las diferencias entre los pares) de $\sigma = 2.9$ y un nivel de significación supuesto de $\alpha = 0.01$. La potencia resultante es de 0.269 y la tasa de error β es de 1 - Potencia = 1 - 0.269 = 0.731. Por lo tanto, este estudio no tendría potencia suficiente, es decir, tiene muy poca potencia. Si ahora se determinara en la planificación cuántas personas se necesitan para medir de forma fiable el efecto aquí descrito – si es que existe – con una potencia suficiente de 0.9, entonces los siguientes resultados

```
> # test for required sample size
> power.t.test(n=NULL, delta=1.1, sd=2.9, sig.level=0.01, power=0.9,
type="paired", alternative="two.sided")
Paired t test power calculation
n = 107
delta = 1.1
sd = 2.9
sig.level = 0.01
power = 0.9
alternative = two.sided
NOTE: n is number of *pairs*, sd is std.dev. of
*differences* within pairs
```

Se necesitan N = 107 personas para poder trabajar seriamente con los mismos parámetros con una potencia objetivo de 0.9 y una tasa de error β de 0.1. Estos cálculos de potencia se utilizan generalmente para calcular a priori el tamaño de la muestra, con el fin de poder clasificar los efectos adecuadamente en una planificación del diseño más apropiada. Es necesario en el caso simple – por ejemplo, la comparación de dos grupos – una distribución no central de la hipótesis alternativa como complemento de la distribución de la hipótesis nula. Esto permite realizar los cálculos correspondientes sobre la base del tamaño del efecto y la tasa de error de tipo I, por ejemplo, para determinar el tamaño de la muestra objetivo o, en función de la pregunta, una cantidad diferente. Mientras que esto puede hacerse con un esfuerzo razonable para una prueba *t*, un análisis de potencia a priori se vuelve muy complicado para los modelos lineales jerárquicos. Un ejemplo de R utilizando `lmer` del paquete R `lme4` es proporcionado por Barstead (2018). El sitio web del software gratuito `MLPowSim` (Browne, Lahi & Parker, 2009) proporciona una visión general de otros paquetes de software para la determinación del tamaño de la muestra. `MLPowSim` permite crear archivos de salida con código R. Un artículo básico sobre cálculos con R y el paquete R `simr`, así como `powerSim()` o `powerCurve()`, procede de Brysbaert y Stevens (2018). El propio paquete R `simr` es introducido por Green y MacLeod (2016). Otras funciones de R para el análisis de potencia de modelos lineales jerárquicos o de efectos mixtos son proporcionadas por los paquetes de R `longpower, pamm` y `sjstats.` Lenth (2018) publicó un applet Java para el análisis de potencia de modelos lineales jerárquicos ("modelos de efectos mixtos").

4.4.3.1 *Ejemplo de una investigación – Simulación de poder – WELL WELL WELL*

Presentamos las funciones más importantes de `simr`. Como conjunto de datos tomamos un estudio parcial del trabajo de Anne A. Huber (2007). En este extenso trabajo sobre la enseñanza y el aprendizaje recíprocos (WELL), se investigó de forma casi experimental la influencia de las formas cooperativas de enseñanza y aprendizaje en las clases de 7º y 8º curso de la Realschule. Para ello, la autora desarrolló un modelo teórico – denominado LUPE – sobre la eficacia de las formas cooperativas de enseñanza y aprendizaje.

El modelo LUPE consta del entorno (de aprendizaje), los procesos (de enseñanza) y las condiciones (iniciales) de aprendizaje. Aquí retomamos un subestudio (ibid., p.153 para una visión general de los subestudios) con las variables *postest de conocimiento* **despuésw**, el *punto temporal* **zeitn** (hubo un total de 6 puntos de medición), el *factor de tratamiento* **gru** con las etapas de tratamiento *trabajo en pareja* vs. *rompecabezas en pareja*, así como *con vs. sin requisitos de aprendizaje* y un *factor de grupo* **grpzugeh**, que indica el grupo de aprendizaje correspondiente. Un gráfico (véase la Fig. 4.41) sugiere que los datos no son lineales a lo largo del tiempo (gráfico reproducido con el amable permiso de la autora). Una de las muchas cuestiones que se plantean es si un modelo debe tener en cuenta los diferentes cambios por persona.

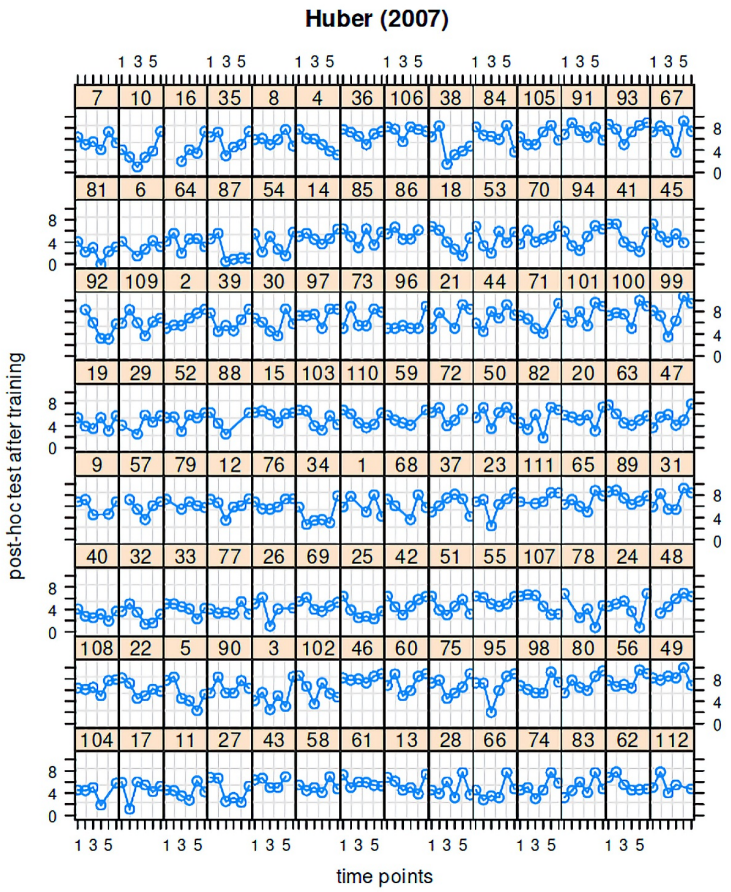


Figura 4.41. *Estudio Huber (2007, evolución temporal – conocimientos después de la intervención)*

```
# plot
daten.i.grp <- groupedData(nachw ~ zeitn | subject, outer=~gru,
inner=~subgr, data=daten.i)
# one person per row
plot(daten.i.grp,layout=c(14,8), main="Huber (2007)",
xlab="time points", ylab="post-hoc test after training")
```

A continuación dejamos de lado el fundamento teórico y las hipótesis derivadas. Los detalles pueden encontrarse en el trabajo de Huber (2007). Una estimación del modelo con `lmer()` del paquete `lme4` de

R permitió aclarar los efectos principales (curso temporal, tratamiento), así como los efectos aleatorios (persona con varios puntos de medición, grupo de aprendizaje 1):

```
# simr
# https://github.com/pitakakariki/simr/issues/96
# on missing fixed effects, test argument, along argument, etc.
m0 <- lmer(nachw ~ poly(zeitn,2) + gru + (1|subject) + (1|grpzugeh),
           data=daten.i)
m1 <- lmer(nachw ~ poly(zeitn,2) + gru + (zeitn-1|subject) +
           (1|grpzugeh), data=daten.i)
m2 <- lmer(nachw ~ poly(zeitn,2) + gru + (zeitn-1|subject) +
           (1|subject) + (1|grpzugeh), data=daten.i)
# fails on R 3.6.3 (Linux, Debian buster, install from CRAN),
# not on R 4.0.2 (Windows)
m3 <- lmer(nachw ~ poly(zeitn,2) + gru + (zeitn|subject) +
           (1|grpzugeh), data=daten.i)
```

La prueba Likelihood-Ratio con `anova()` apoya estadísticamente de forma clásica la inclusión del efecto aleatorio **zeitn**, es decir, la estimación de diferentes pendientes y niveles de regresión por persona.

```
> anova(m0,m1,m2,m3)
refitting model(s) with ML (instead of REML)
Data: daten.i
Models:
m0: nachw ~ poly(zeitn, 2) + gru + (1 | subject) + (1 | grpzugeh)
m1: nachw ~ poly(zeitn, 2) + gru + (zeitn - 1 | subject) +
           m1: (1 | grpzugeh)
m2: nachw ~ poly(zeitn, 2) + gru + (zeitn - 1 | subject) +
           m2: (1 | subject) + (1 | grpzugeh)
m3: nachw ~ poly(zeitn, 2) + gru + (zeitn | subject) + (1 | grpzugeh)
   npar AIC  BIC   logLik deviance Chisq  Df Pr(>Chisq)
m0  9   2475 2516 -1229   2457
m1  9   2477 2517 -1229   2459      0.00  0
m2 10   2465 2510 -1223   2445     13.21  1  0.00028 ***
m3 11   2467 2516 -1223   2445      0.23  1  0.62788
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Con do`Test()` de `simr`, se pueden probar tanto los efectos fijos como los aleatorios.

```
# nachw ~ poly(zeitn,2) + gru + (zeitn-1|subject) +
# (1|subject) + (1|grpzugeh)
> doTest(m2, fcompare(~ poly(zeitn,2)))
p-value for model comparison: 0.158
--------------------
Test: Likelihood ratio
Comparison to ~poly(zeitn, 2) + [re]
> doTest(m2, rcompare(~ (1|subject)))
p-value for model comparison: 0.00141
--------------------
Test: Likelihood ratio
Comparison to [fe] + ~(1 | subject)
> doTest(m2, rcompare(~ (1|grpzugeh)))
p-value for model comparison: 8.5e-27
--------------------
Test: Likelihood ratio
Comparison to [fe] + ~(1 | grpzugeh)
> doTest(m2, rcompare(~ (zeitn-1|subject)))
p-value for model comparison: 0.00017
--------------------
Test: Likelihood ratio
Comparison to [fe] + ~(zeitn - 1 | subject)
```

con diversas pruebas, véase ?tests. Las pruebas satisfactorias pueden transferirse directamente a la simulación de potencia mediante powerSim(),

```
# simulate test to leave out unique level for each subject
p1 <- powerSim(m2, test=rcompare(~ (1|subject)), nsim=50)
```

se recibe el output

```
> print(p1)
Power for model comparison, (95% confidence interval):
88.00% (75.69, 95.47)
Test: Likelihood ratio
Comparison to [fe] + ~(1 | subject)
Based on 50 simulations, (1 warning, 0 errors)
alpha = 0.05, nrow = 672
Time elapsed: 0 h 0 m 6 s
nb: result might be an observed power calculation
```

y, a continuación, puede simular la potencia para diferentes valores de las variables de interés, que luego se pueden mostrar gráficamente. Por razones de tiempo de cálculo, limitamos el número de simulaciones a nsim = 50. Por defecto, `powerSim()` tiene nsim = 1000. Además, el nivel de significación se establece por defecto con el parámetro alfa = 0.05. La figura 4.42 contiene el gráfico.

```
pc1 <- powerCurve(m2, along="zeitn",
test=rcompare(~ (1|subject)), nsim=50)
plot(pc1)
```

(Una advertencia necesaria cuando se experimenta con R: Puede observarse que, dependiendo de la versión de R, algunos de estos modelos pueden estimarse y otros no. En el caso, algunos modelos no pudieron estimarse con R 3.6.3 (Linux/CRAN), pero sí con R 4.0.2 (Windows/CRAN). Esto debe tenerse en cuenta al ejecutar los scripts de R.)

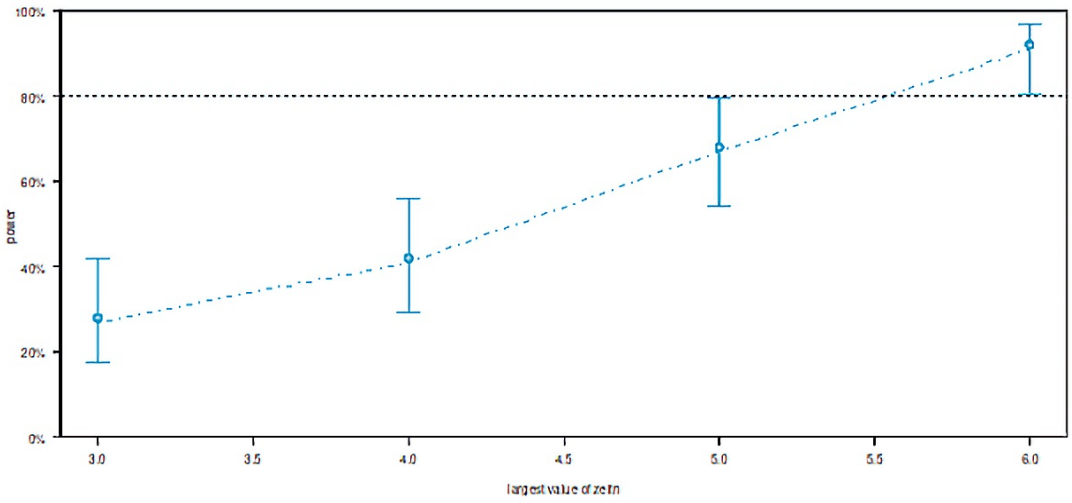


Figura 4.42. *Estudio Huber (2007, simulación potencia observada "dejar fuera nivel único por sujeto")*

Se puede demostrar que los cambios en el tamaño de la muestra, aquí realizado a través del número de puntos de medición zeitn, conducen a cambios significativos en la simulación de potencia, como era de esperar. Empecemos probando el efecto aleatorio zeitn como una pendiente aleatoria. Aquí examinamos la cuestión de si tiene sentido utilizar una línea de regresión diferente con su propia pendiente para cada persona, en lugar de suponer simplemente un nivel diferente (intercepto aleatorio). Esta simulación pertenece al dominio de la potencia observada (= post-hoc) y puede utilizarse para planificar un estudio futuro. Debido

a las fluctuaciones aleatorias de las muestras, no es aconsejable hacer observaciones retrospectivas demasiado fuertes sobre los resultados empíricos sin introducir cambios específicos. Los posibles cambios – véase más adelante – son, por ejemplo, aumentar o disminuir el tamaño de la muestra o especificar un determinado tamaño del efecto y ajustar la simulación de la potencia. Por ejemplo, se puede seleccionar un efecto más pequeño en función de los datos empíricos y la potencia simulada. Esto tendría el sentido estimar la constancia del efecto encontrado, ya que los efectos más pequeños sólo se encuentran con potencia suficiente, mientras que los efectos más grandes son más fáciles de identificar. Comenzamos examinando la potencia de los puntos de medición en función del número de puntos de medición (véase la Fig. 4.43).

```
# we test in dependence of the measurement points on 'zeitn'
# (total of n=6)
pc1.1 <- powerCurve(m2, along="zeitn", test=rcompare(~ (1|subject)),
nsim=50, breaks=1:6)
print(pc1.1)
plot(pc1.1)
```

La figura 4.43 (arriba) visualiza el resultado incluyendo los siguientes cambios. Ahora reducimos globalmente el número de puntos de medición, pero sólo consideramos la variable zeitn y no zeitn dentro de sujeto. Lo que hacemos es considerar la variable zeitn sin su incrustación en la variable sujeto y tratar globalmente zeitn como una variable de agrupación con k niveles de factor. En la parte derecha de la salida de R se encuentra el número de filas. Esto da información sobre la base de datos de la simulación de potencia. Por lo tanto, el resto del código R indica el número de niveles de factores en los que se basa la simulación de potencia.

```
# wrong call, see manpage ?extend
# reduce number of levels and see how power drops down into the cellar tp.a <- 6
# factor levels
6*(1:6)
# extend model by this new configuration
m2.a <- extend(m2, within="zeitn", n=tp.a)
pc2.a <- powerCurve(m2.a, along="zeitn", test=rcompare(~ (1|subject)),
nsim=50, breaks=1:tp.a)
print(pc2.a)
plot(pc2.a)
```

Obviamente, el poder se va por el desagüe (véase la Fig. 4.43, abajo). El código R anterior no es especialmente inteligente y sólo tiene fines demostrativos. Si busca en el manual de `extend()` con la orden `?extend`, se dará cuenta de que existen las variables *`along`* y *`within`*, que no deben confundirse. La explicación en la página de ayuda dice exactamente

```
along: the name of an explanatory variable. This variable will
have its number of levels extended. within: names of grouping
variables, separated by "+" or ",". Each combination of groups
will be extended to 'n' rows.
```

Por lo tanto, la llamada anterior se especificó incorrectamente. Con *`along`* se aumenta el número de niveles de una variable explicativa (covariables), mientras que con *`within`* se cambia el número de combinaciones (niveles de factor) de las variables de agrupación.

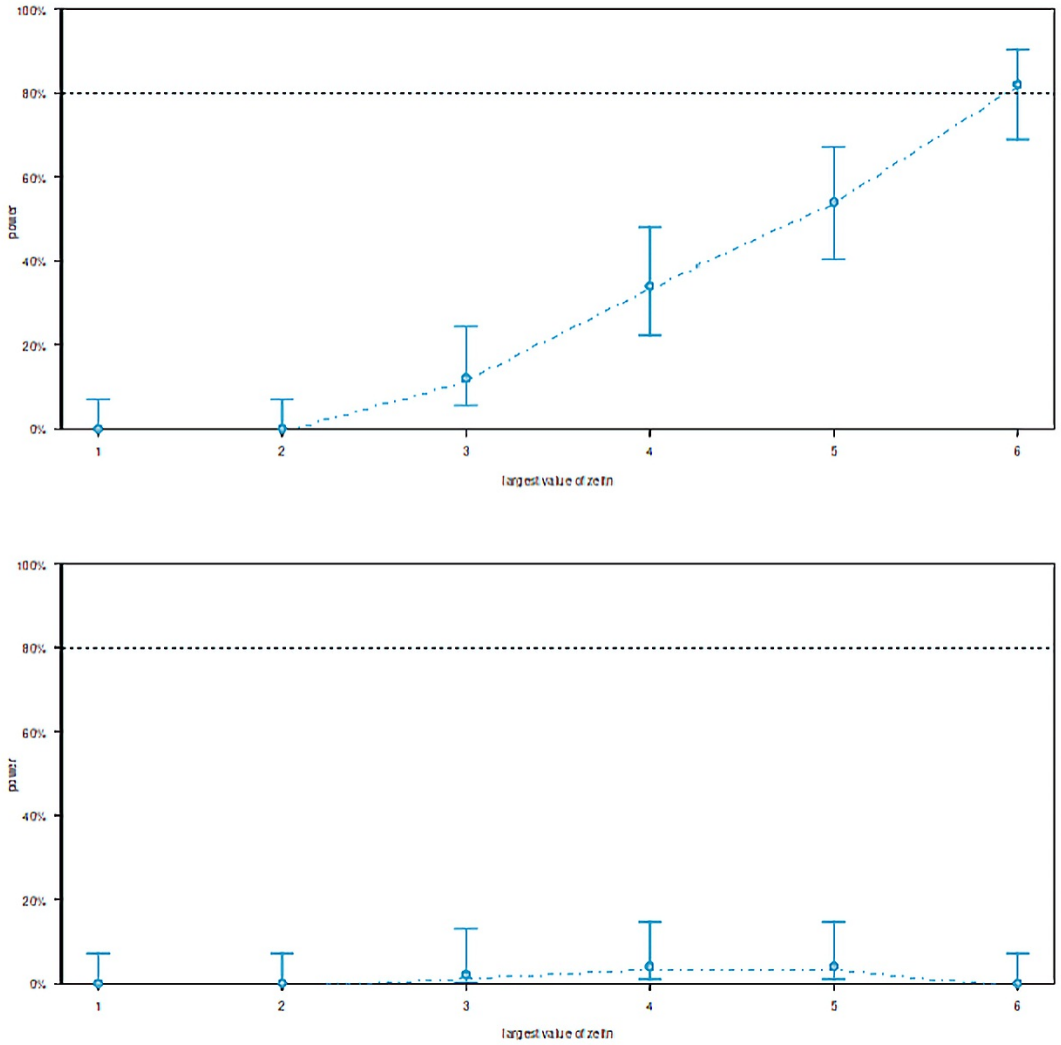


Figura 4.43. *Estudio Huber (2007, simulación de la potencia observada y dependencia de zeitn)*

En lugar de utilizar sólo *within* con `zeitn` y sin incorporar una variable de agrupación como `subject`, lo correcto habría sido utilizar *along*. Es mejor aumentar sólo el número de veces de medición:

```
tp.b <- 25
# factor levels
length(levels(daten.i$subject))
112*(1:25)
# increase 'within' - wrong!
m2.b <- extend(m2, within="zeitn", n=tp.b)
pc2.b <- powerCurve(m2.b, along="zeitn", test=rcompare(~ (1|subject)),
           nsim=50, breaks=1:tp.b)
print(pc2.b)
plot(pc2.b)
# DO IT BETTER!
# now the proper call
tp.c <- 25
# factor levels
length(levels(daten.i$subject))
112*(1:25)
# increase 'along' - correct!
m2.c <- extend(m2, along="zeitn", n=tp.c)
pc2.c <- powerCurve(m2.c, along="zeitn", test=rcompare(~ (1|subject)),
```

```
            nsim=50, breaks=1:tp.c)
print(pc2.c)
plot(pc2.c)
```

Esto se ve mucho mejor en el gráfico (véase la Fig. 4.44, más abajo) y de la forma que esperábamos. Muestra que no merece la pena invertir en más de seis puntos temporales en el contexto de todas las demás variables (tamaño de la muestra, covariables). De lo contrario, el estudio está *sobredimensionado* y esto no nos dice mucho más y, en cambio, conduce a inversiones innecesarias si pensamos en un *cálculo de costes completo* (véase el capítulo 7.10.2 o 4.3.3.5). Ahora podemos intentar considerar la variable `zeitn` dentro del tema y dejar constante el número original de puntos de medición de seis. Es decir, ahora hay seis puntos de medición por `sujeto persona` dentro de `zeitn`. De este modo se examina el número de combinaciones de nivel de factor presentes para probar el efecto aleatoria `zeitn` *dentro del* `sujeto`. Comenzamos de nuevo calculando el número de combinaciones de nivel de factor.

```
# increase levels again and see power rising high
tp.d <- 6
dim(daten.i)
# factor levels
length(daten.i$subject) * (1:tp.d) 672*(1:6)
```

Y ahora la simulación:

```
# increase 'within' - correct!
m2.d <- extend(m2, within="zeitn+subject", n=tp.d)
pc2.d <- powerCurve(m2.d, along="zeitn", test=rcompare(~ (1|subject)),
            nsim=50, breaks=1:tp.d)
print(pc2.d)
plot(pc2.d)
# here large power is already with zeitn >=3 (power=.7)
# and zeitn >=4 (power >0.95)
```

La potencia (véase la Fig. 4.44) parece ahora claramente diferente en función del número de puntos de medición por persona y, en consecuencia, tiene sentido utilizar ecuaciones de regresión separadas con intercepción y pendiente en el eje Y para cada persona. Esto demuestra que el uso estándar de modelos lineales jerárquicos, como sugieren Gelman y Hill (2007), es probablemente rentable en la mayoría de los casos. Como se señaló al principio, se pueden especificar nuevos tamaños del efecto aparte de la potencia observada "post-hoc" clasificada críticamente, con el fin de derivar simulaciones para futuros estudios.

Utilizamos la simulación para examinar dos tamaños del efecto que hemos especificado, uno mayor y otro menor que el efecto medido empíricamente. Por razones de simplificación, elegimos un modelo sencillo sin pretender que – véanse los argumentos y explicaciones anteriores – sea también el mejor modelo.

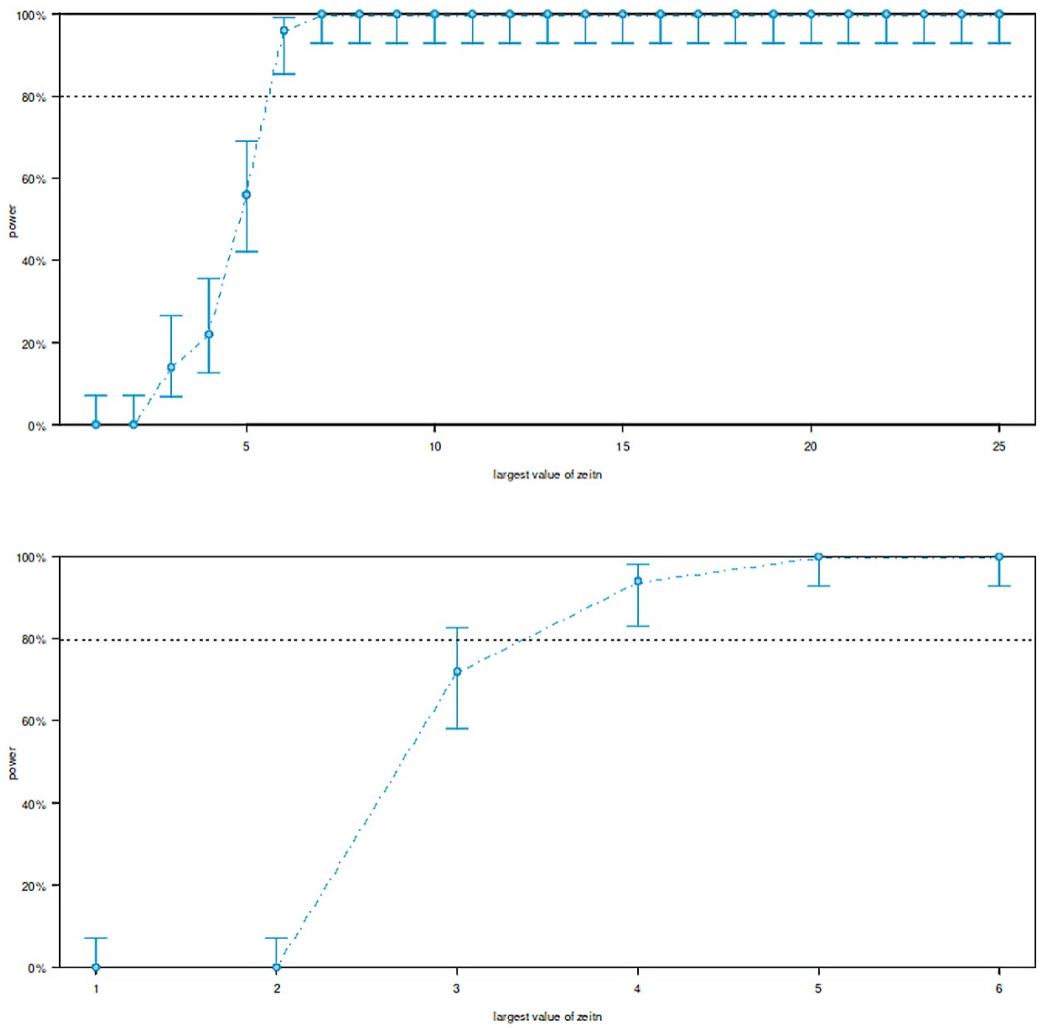


Figura 4.44. *Estudio Huber (2007, simulación de la potencia observada en función del número de niveles)*

```
# test for a specific effect against a value
# we use a simple model out of didactic reasons
# with only one fixed effect and one random effect mx <- lmer(nachw ~ gru +
    (1|subject),
# data=daten.i)
summary(mx)
fixef(mx)
mx1 <- mx
levels(daten.i[,"gru"]) contrasts(daten.i[,"gru"])
# see second column
# we test 'PAm+PPm' vs. 'PAo+PPo' fixef(mx1)["gru2"]
fixef(mx1)["gru2"] <- 0.8
fixef(mx1)["gru2"]
mx2 <- mx
fixef(mx2)["gru2"]
fixef(mx2)["gru2"] <- 0.5
fixef(mx2)["gru2"]
ps.mx <- powerSim(mx, nsim=50)
pc.mx <- powerCurve(mx, nsim=50)
ps.mx1 <- powerSim(mx1, nsim=50)
pc.mx1 <- powerCurve(mx1, nsim=50)
ps.mx2 <- powerSim(mx2, nsim=50)
```

```
pc.mx2 <- powerCurve(mx2, nsim=50)

# be aware that the 'lower' estimate is detected more difficult than
# the 'larger' estimate i.e. different power with growing number of
# levels in gru (= group)
# this is congruent with an assumption that large effects are detected # more easily
than # small effects
# outputs
print(ps.mx)
print(ps.mx1)
print(ps.mx2)
print(pc.mx)
print(pc.mx1)
print(pc.mx2)
# plots
plot(pc.mx)
plot(pc.mx1)
plot(pc.mx2)
```

Obviamente, la potencia de un efecto menor es menor que la de redondear un efecto mayor. Esto es lo esperado, porque los efectos mayores son más fáciles y rápidos de identificar. Para redondear efectos menores, necesitamos una muestra mayor, más puntos de medición, una división diferente de los grupos, etc., dependiendo del contexto.

Con un tamaño del efecto dado, que no se estima a partir de datos empíricos, podríamos empezar desde cero utilizando `powerSim()`, `powerCurve()`, `extend()`, etc. Así, se puede definir el campo de investigación de un estudio en términos de la gama de efectos, cómo se relacionan y qué papel desempeña en ello el tamaño de la muestra (operacionalizado como el número de personas, los tiempos de medición, etc.). Se trata de un ámbito complejo y, por lo tanto, no se pueden hacer afirmaciones generales sobre el tamaño de los efectos y la potencia que no sean adecuadas al caso. Sin embargo, en el caso del estudio de Anne A. Huber (2007) mencionado anteriormente, puede verse que la opción con seis momentos de medición diferentes en el contexto de todas las variables se eligió intuitivamente de forma muy favorable para encontrar también los efectos presentes en los datos con un esfuerzo justificable. "Intuitivamente" porque no había sido posible realizar estudios piloto exhaustivos en varios puntos temporales de medición en el período previo al estudio principal. Una futura repetición podría basarse en esto y ajustar el diseño en consecuencia. Del mismo modo, ahora habría que repetir todo el procedimiento para cada variable adicional, de modo que pueda discutirse una evaluación exhaustiva de los tamaños de los efectos en el contexto de la potencia.

Además de los artículos sobre `simr` mencionados anteriormente, las viñetas del paquete R (Green, MacLeod & Alday, 2019-01-29) y las discusiones en la página del desarrollador de `simr()` (Green, 2018-01-24) son muy útiles para aprender a utilizar las funciones correctamente. Especialmente en las viñetas, se muestra adicionalmente cómo trabajar con un estudio piloto existente o datos empíricos por un lado, pero igualmente cómo, basándose en variables dadas y asumidas, se puede construir una simulación completa para planificar cuidadosamente un estudio sin piloto en absoluto. El paquete R `simr`, sin embargo, permite muchas más posibilidades de las que se muestran aquí. Con `?modify`, por ejemplo, se puede cambiar cualquier área de los objetos `lmer` para adaptarlos a los análisis de potencia. Con `?tests`, están disponibles diversas variantes de pruebas (incluyendo bootstrap paramétrico) y los propios datos se pueden modificar con `?getData`.

Alternativamente, consideramos el mismo conjunto de datos con `lmmpower()` del paquete R `longpower`, que permite simular la potencia para datos a largo plazo. Consideramos el mismo modelo que el anterior.

```
# take  a reduced model
m0.red  <- lmer(nachw ~ poly(zeitn, 2) + gru + (1+zeitn|subject),
data=daten.i)
summary(m0.red)
```

Con `lmmpower()`, al igual que con otras funciones comparables en R, se pueden hacer especificaciones para los parámetros (tamaño del efecto, potencia, tamaño de la muestra, tasa de error de tipo I, dirección de

la hipótesis, etc.), por lo que un parámetro debe pasarse con cero, ya que luego se estima. Con `lmmpower()`, también se añade información sobre los puntos temporales *t* y las estimaciones piloto de las varianzas de la selección aleatoria y la varianza residual, respectivamente. Esto permite planificar un estudio de forma concreta, independientemente de los datos empíricos y de los supuestos dados sobre los efectos. La función espera que las estimaciones piloto procedan de grupos "placebo" adecuados (= efectos nulos) y los parámetros de interés se formulan como tasas de cambio a lo largo del tiempo relacionadas con la variable dependiente. Se distingue entre `delta` y `pct.change`. El primero describe el cambio desde la estimación piloto del parámetro de interés y se calcula mediante el segundo si falta información. `pct.change` representa el cambio porcentual del parámetro de interés – extraído de la estimación piloto (placebo o efecto nulo).

```
> lmmpower(m0.red, n=NULL, parameter=6, delta=-0.2,
+ t=seq(1,6,1), power=0.8)
Longitudinal linear model slope power calculation
(Diggle et al 2002, page 29)
N = 111.1344
n = 55.56718, 55.56718
delta = -0.2
sigma2 = 1
sig.level = 0.05
power = 0.8
alternative = two.sided
delta.CI = -2.183158, 1.783158
gru3 = -0.04392088
gru3 CI = -0.4794310, 0.3915893
n.CI = 0.6990349, 0.6990349, 0.4663459, 0.4663459
NOTE: N is *total* sample size and n is sample size in *each* group
R:
     [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
[1,] 2.6234695 0.7312596 0.7947067 0.8581537 0.9216007 0.9850478
[2,] 0.7312596 2.7802044 0.9178353 1.0111232 1.1044110 1.1976989
[3,] 0.7947067 0.9178353 2.9966209 1.1640927 1.2872213 1.4103500
[4,] 0.8581537 1.0111232 1.1640927 3.2727191 1.4700316 1.6230011
[5,] 0.9216007 1.1044110 1.2872213 1.4700316 3.6084989 1.8356522
[6,] 0.9850478 1.1976989 1.4103500 1.6230011 1.8356522 4.0039603
> lmmpower(m0.red, n=NULL, parameter=6, pct.change=0.2,
+ t=seq(1,6,1), power=0.8)
Longitudinal linear model slope power calculation
(Diggle et al 2002, page 29)
N = 57611.12
n = 28805.56, 28805.56
delta = -0.008784176
sigma2 = 1
sig.level = 0.05
power = 0.8
alternative = two.sided
delta.CI = -0.09588621, 0.07831785
gru3 = -0.04392088
gru3 CI = -0.4794310, 0.3915893
n.CI = 362.3738, 362.3738, 241.7498, 241.7498
NOTE: N is *total* sample size and n is sample size in *each* group
R:
     [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
[1,] 2.6234695 0.7312596 0.7947067 0.8581537 0.9216007 0.9850478
[2,] 0.7312596 2.7802044 0.9178353 1.0111232 1.1044110 1.1976989
[3,] 0.7947067 0.9178353 2.9966209 1.1640927 1.2872213 1.4103500
[4,] 0.8581537 1.0111232 1.1640927 3.2727191 1.4700316 1.6230011
[5,] 0.9216007 1.1044110 1.2872213 1.4700316 3.6084989 1.8356522
[6,] 0.9850478 1.1976989 1.4103500 1.6230011 1.8356522 4.0039603
```

Esto se puede trazar como de costumbre sobre diferentes valores de `delta` para calcular el tamaño de muestra necesario relacionado con los tamaños de los efectos. Repetimos lo mismo para el cambio porcentual `pct.change` en lugar de para una secuencia de tamaños del efecto.

```
# analyse for n based on delta sequence and percentage change R-Code
ds <- seq(-0.4,-0.01,0.005)
ds.l <- length(ds)
pct <- seq(0.1,0.4, length.out=ds.l)
m0.red.lmmpwr <- matrix(data=NA, nrow=ds.l, ncol=2)
colnames(m0.red.lmmpwr) <- c("N(delta)","N(pct)")
for(i in 1:ds.l)
{
m0.red.lmmpwr[i,"N(delta)"] <- lmmpower(m0.red, n=NULL, parameter=6,
           delta=ds[i], t=seq(1,6,1), power=0.8)[["N"]]
m0.red.lmmpwr[i,"N(pct)"] <- lmmpower(m0.red, n=NULL, parameter=6,
           pct.change=pct[i], t=seq(1,6,1), power=0.8)[["N"]]
# N = 2*n (output of lmpower) = total sample size
# n = sample size for each group
}
head(m0.red.lmmpwr)
dim(m0.red.lmmpwr)
```

En el gráfico (véase la Fig. 4.45), el tamaño de la muestra se traza en la escala `log()` para diferentes tamaños del efecto con el fin de facilitar su lectura. Como era de esperar, el tamaño de muestra necesario aumenta masivamente a medida que los tamaños del efecto se hacen más pequeños.

```
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l", mfrow=c(2,1)) R-Code
plot(ds, log(n1), main="", pre.plot=grid(), bty="l",
xlab=expression(delta), ylab="log(n)", col="magenta", type="l")
plot(pct, n2, main="", pre.plot=grid(), bty="l", xlab="pct change",
           ylab="n", col="magenta", type="l")
mtext("Huber (2007)", outer=TRUE, line=-1.7, cex=1.5, side=3)
mtext("Simulation of power with lmmpower()", outer=TRUE, line=-3.1,
           cex=1, side=3)
```

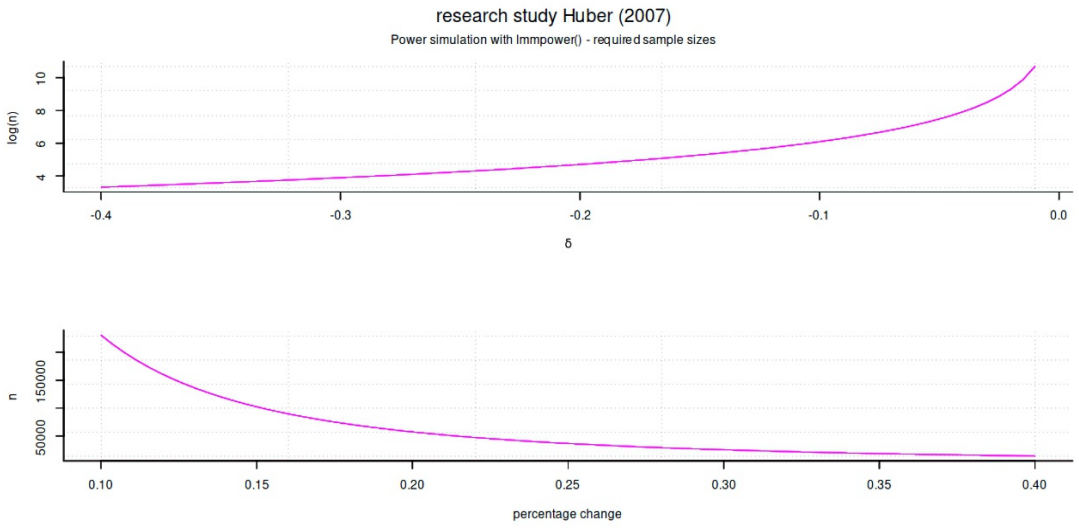


Figura 4.45. *Estudio Huber (2007, simulación con `lmmpower()`, tamaño de muestra requerido).*

Esta pequeña idea debería bastar. Los detalles pueden encontrarse en las viñetas de `lmmpower` (Magnusson, 2018-08-14).

Como señalan no inesperadamente Sedlmeier y Gigerenzer (1989) y Sterling, Rosenbaum y Weinkam (1995), la potencia típica de los estudios psicológicos no es particularmente alta e incluso ha disminuido desde el conocido estudio de Cohen (1962), que Sedlmeier y Gigerenzer (1989) replican para el período 1960 a 1984 para el *Journal of Abnormal Psychology*. Sedlmeier y Gigerenzer (1989) se preguntan por tanto, en vista de la escasa potencia y la frecuente uso de procedimientos de ajuste para α (ibid., p.313),

> „Must we conclude that researchers stubbornly neglect a major methodological issue over decades? Or should we assume that they are intuitively right and that we really do not need more power than .37?“

Otros autores asumen una potencia típica del 60%, mientras que la tasa de resultados significativos supera el 90% (Sterling, 1959; Sterling, Rosenbaum y Weinkam, 1995). Esta discrepancia se explica de diferentes maneras. Por un lado, puede tratarse de falsos positivos; por otro, refleja un enfoque rígido en los resultados significativos, independientemente de la potencia real. Asimismo, entran en juego las mediciones inflacionistas del tamaño del efecto. Todo esto junto, Schimmack (2012) apunta a una estimación sesgada de la potencia real de los estudios. El autor (2016a) es especialmente crítico en este punto:

> „If the actual power is less than 50 %, it means that a typical study in psychology has a larger probability to fail (produce a false negative result) than to succeed (rejecting a false nullhypothesis). Conducting such low powered studies is extremely wasteful. Moreover, few researchers have resources to discard 50 % of their empirical output. As a result, the incentive for the use of questionable research practices that inflate effect sizes is strong.“

Como caso ejemplar de tal sesgo se cita el mencionado estudio de Bem (2011a, véase el capítulo 4.4.2.2). En él, 9 de cada 10 estudios se notifican como significativos al nivel convencional del 5 %. Las posibilidades de que esto se base en un error tipo I común (= falsos positivos) son pequeñas. Son peores que ganar la lotería o que te caiga un rayo con

```
> # check via binomial test
> # p-value = probability of empirical value or extreme!
> binom.test(x=9, n=10, p=.05)
Exact binomial test
data: 9 and 10
number of successes = 9, number of trials = 10, p-value = 1.865e-11
alternative hypothesis: true probability of success is not equal to 0.05
95 percent confidence interval:
0.5549839 0.9974714
sample estimates:
probability of success
0.9
```

En concreto, existe una probabilidad de 1 : $5.36e^{+10}$. Esto corresponde al valor p de la prueba binomial, ya sea calculando manualmente la probabilidad de 9 o 10 sucesos con éxito. cálculo de

```
> # exact probability for 9 of 10 with p=0.05
> 1/( choose(10,9) * 0.05^9 * (1-0.05)^(10-9) )
[1] 53894736842
> # exact probability for 10 out of 10 with p=0.05
> 1/( choose(10,10) * 0.05^10 * (1-0.05)^(10-10) )
[1] 1.024e+13
> # the same, but we have to sum up for the p-value
> 1/(dbinom(x=9, prob=0.05, size=10) + dbinom(x=10, prob=0.05, size=10))
[1] 53612565445
> # equals
> 1/ sum(dbinom(x=9:10, prob=0.05, size=10))
[1] 53612565445
> # equals
> 1/binom.test(x=9, n=10, p=.05, alternative="greater")$p.value
[1] 53612565445
> 1/binom.test(x=9, n=10, p=.05, alternative="less")$p.value
[1] 1
> 1/binom.test(x=9, n=10, p=.05, alternative="two.sided")$p.value
[1] 53612565445
```

o la distribución de densidad de la distribución binomial con `dbinom()`. Dado que el valor *p* se define para cubrir el rango de todos los valores (empíricos o valores más extremos), no basta con calcular sólo el

caso de 9 aciertos. También hay que añadir el caso de 10 aciertos potenciales al cálculo. La suma de estas dos probabilidades arroja el valor *p* de `binom.test()`.

Volvamos a la cuestión de si el estudio de Bem es serio y realista. En cualquier caso, su resultado parece "extraordinario", y no parece deberse a que los estudios no significativos no se mencionaron, porque eso serían

```
> # number of research studies required
> # to get 9 'significant' on 5% crit. level
> 100/5*9
[1] 180
> # number of participants required if one study has n=100
> 100/5*9 * 100
[1] 18000
```

Después de todo, normalmente se necesitarían 180 estudios para obtener 9 estudios estadísticamente significativos al nivel del 5%, ya que sólo cabe esperar que uno de cada 20 estudios sea estadísticamente significativo. Dado que Bem cita un tamaño de muestra de 100, habría necesitado 100 * 180 = 18 000 sujetos de estudio. Esto parece poco realista. A este respecto, el debate se centra naturalmente a la cuestión de la potencia. Hay que diferenciar entre poder real y poder observado. En el contexto de la teoría Neyman-Pearson, la potencia observada debe converger a la potencia verdadera a largo plazo y estimarla de forma fiable si no se conoce. Sin embargo, esto sólo es cierto si no existen sesgos sistemáticos. Técnicamente, la potencia puede determinarse por el número de resultados significativos de los estudios repetidos. Teóricamente, una subestimación de la potencia real sería tan probable como una sobreestimación, siempre que el número de estudios sea pequeño. Sin embargo, las revistas informan con frecuencia de resultados significativos, lo que sugiere un sesgo sistemático – por ejemplo, sesgo de publicación (Carroll, Toumpakari, Johnson & Betts, 2017; Mlinaric, Horvat & Smolcic, 2017; Murad, Chu, Lin & Wang, 2018). Así, el número de resultados significativos es un mal estimador de la verdadera potencia de los estudios. Schimmack (2012) propone un índice de incredibilidad (= índice i) para estimar la frecuencia de resultados significativos. En el caso de Bem, se calcula de la siguiente manera para una potencia supuesta del 60% y 9 de cada 10 éxitos: En primer lugar, se requiere la probabilidad binomial acumulativa $p(X \geq x)$ si x = número de aciertos, aquí x = 9. Según el cálculo resulta

```
> # calculate probabilities
> # Bem study
> # assumed power=0.6 -> get 9 out of 10 successful
> 1-pbinom(9,10,.6) + dbinom(9,10,.6)
[1] 0.0463574
> 1-cumsum(dbinom(0:(10-1),10,.6))
[1] 0.999895142 0.998322278 0.987705446 0.945238118 0.833761382
[6] 0.633103258 0.382280602 0.167289754 0.046357402 0.006046618
```

El resultado es una probabilidad de p = 0.046, es decir, 4.6%. El *índice i sugerido i = 1- p* indica que una serie de estudios parece poco fiable porque debería haber producido más resultados no significativos esperados. En el caso que nos ocupa, el índice i = 1 - 0.046 = 0.954. Implícitamente, esto indica una selección que debe especificarse con mayor precisión, de modo que los resultados comunicados sugieren pruebas más sólidas de una hipótesis de lo que probablemente habría sido el caso si se hubiera utilizado una serie completa sin selección.

Una crítica al índice i es que se basa en la potencia post hoc observada. La potencia post hoc de los estudios individuales se ve afectada por el hecho de que los intervalos de confianza son demasiado amplios y, por tanto, cubren por lo que abarcan casi todo el espacio de posibilidades y, por tanto, son bastante poco informativos, lo que no debe confundirse con sesgado. La incertidumbre respecto a una medición concreta es muy elevada. Esto ya no es así en los metaanálisis que examinan estudios múltiples y no singulares, porque siguen el enfoque Neyman-Pearson de estudios repetidos y, por tanto, la verdadera potencia puede calcularse aproximadamente con un número creciente de estudios. Los estudios de potencia post-hoc también pueden generarse mediante simulaciones utilizando un único conjunto de datos (véase el capítulo 4.4.3.1 para un ejemplo de investigación empírica). La cuestión entonces es qué parte del espacio de posibilidades del objeto estudiado son capaces de cubrir y cuánto ruido ha introducido la simulación. Más importante es la objeción

de que las probabilidades no hacen afirmaciones sobre el tamaño de los efectos y el número de estudios investigados. Además cuantas más pruebas enumere una revista y más artículos contenga, el metaanálisis con el índice *i* conduce automáticamente a un número mayor, incluso si la potencia es alta por término medio y el media es alta y los estudios son reputados e imparciales. Todo esto requiere una mejora, propuesto aquí como el índice *R* según Schimmack (2016a). Esto pertenece al ámbito de las *réplicas*.

4.4.4 Replicaciones

De los comentarios sobre la potencia sólo queda un paso para llegar al tema de las *réplicas*. Las réplicas no sólo son importantes en el contexto de la teoría de Neyman-Pearson, que se basa en el concepto de medición repetida y hace afirmaciones sobre exactamente eso: réplicas. Más bien, la replicación es uno de los instrumentos de investigación más importantes de todos, ya que utiliza la tesis del *muestreo aleatorio* para apuntalar la constancia o robustez de los efectos y, por tanto, la prueba empírica o validez de las hipótesis y (elementos de) la teoría. Los estudios rara vez pueden reproducirse directamente, por lo que, en el caso de los efectos constantes, puede suponerse cierta robustez de los efectos de interés en vista de diversos efectos específicos del estudio. Sin embargo, hay pruebas de que las réplicas fracasan o sólo muestran efectos reducidos o incluso magnificados. Así pues, existen variaciones en ambas direcciones entre las réplicas en cuanto a sus efectos, lo cual es de esperar estadísticamente (por ejemplo, Open Science Collaboration, 2015; Klein et al., 2014). Sin embargo, muchos estudios no replicados sí entran con el tiempo en los libros de texto o en el repertorio teórico común de una profesión sin señalarlo explícitamente. Lo mismo ocurre con los estudios que nunca llegan a publicarse en una revista porque no eran significativos por varios motivos, lo que no es en absoluto sinónimo de científicamente irrelevantes.

Como señalan los autores de Open Science Collaboration, también faltan normas claras para la replicación, lo que se complica por el hecho de que no se dispone de toda la información necesaria para replicar y sopesar los efectos. La replicación implica la repetición exacta de un estudio empírico en las mismas condiciones para comprobar si los efectos registrados pueden repetirse, es decir, replicarse (o no). La idea de la replicación siempre implica dos tendencias contradictorias que sólo pueden resolverse hasta cierto punto; y en cada caso individual debe decidirse cuál parece más importante. Esta contradicción se basa en la *identidad de replicación* frente a *aprendizaje de la experiencia*.

- Por un lado, una réplica exacta e idéntica de un estudio conduce desgraciadamente a que obviamente no se haya aprendido nada del primer estudio para hacerlo mejor en el futuro. El empirismo debe estimular el aprendizaje. Si ahora tiene lugar un proceso de aprendizaje, el estudio en cuestión ya no puede repetirse exactamente porque esto significaría no utilizar lo aprendido. En cambio, los resultados del primer estudio sí pueden examinarse críticamente.
- Por otra parte, si un estudio no se repite exactamente, sino teniendo en cuenta el proceso de aprendizaje descrito, los resultados no pueden compararse de forma idéntica entre los dos estudios, porque puede que ahora se trate de preguntas y afirmaciones ligeramente diferentes. Esto no significa que se esté realizando una comparación de manzanas con peras, sino quizás se compare manzanas de variedades tan diferentes que cualquier diferencia sólo se puede atribuir condicionalmente al tratamiento (en el caso de la investigación experimental) y posiblemente a las diferencias de diseño entre los estudios o al hecho que los estudios representan diferentes preguntas de investigación.
- Una solución a este dilema principal puede encontrarse en proyectar más ampliamente la réplica o el metaanálisis en términos de la pregunta de investigación y valorar mejor los resultados equivalentes, ya que tal vez la probabilidad de obtener resultados teóricamente igual orientados y confirmados sea menor si los estudios adoptan perspectivas diferentes sobre el mismo fenómeno.

Ser consciente de la *necesidad esencial de replicación* forma parte hoy en día de la corriente principal de la investigación. Por este motivo, diversos autores e investigadores se han unido para replicar estudios ya publicados bajo el término de *replicación pre-registrada* (por ejemplo, Lei, Gelman & Ghitza, 2017). A veces, esto implica incluso la cooperación con los investigadores de los estudios originales para garantizar la

comparabilidad de las réplicas y poder documentar e informar de cualquier diferencia (Simons, Holcombe & Spellman, 2014). En la actualidad existen directrices sobre cómo llevar a cabo dichas réplicas (APS, s.f.). En 2013, un grupo más amplio de investigadores estableció el proyecto "Many Labs Replication Project" (s.f.), que intentó replicar 13 efectos en psicología, con la participación de 36 muestras y más de 6 000 individuos (Klein et al., 2014).

4.4.4.1 Índice R

Para los metaanálisis, Schimmack (2016a) propone un índice R(eplicación) cuantitativo (Schimmack, 2018a) para evaluar la integridad de los estudios. Al hacerlo, el índice R debería ayudar a identificar y exponer prácticas de investigación cuestionables.

El índice R toma valores entre cero y uno. No debe calcularse para estudios individuales, sino siempre para un conjunto de estudios, ya que (véase más arriba) la potencia observada de estudios individuales suele dar una visión poco clara de la potencia real. El índice R conoce los siguientes elementos

$$\text{Tasa de inflación} = \text{Tasa de éxito} - \text{Mediana (potencia observada)} \quad (4.23)$$

$$\begin{aligned}\text{Índice R} &= \text{Mediana (potencia observada)} - \text{Tasa de inflación} \quad (4.24)\\ &= 2 * \text{Mediana (potencia observada)} - \text{Tasa de éxito}\end{aligned}$$

El índice R no es una medida de potencia, pero tiene una relación monótona con la potencia. Cuanto mayor sea el índice R, más cerca estarán los estudios publicados y examinados meta-analíticamente de la potencia real y del porcentaje de éxito real. El índice R es una medida descriptiva que no se basa únicamente en la significación, sino que dice poco específicamente sobre la causalidad (por ejemplo, p-hacking, sesgo selectivo de publicación, es decir, sólo se publican los estudios altamente significativos) y sólo insinúa posibles incoherencias. "Así que se forma un tamaño del efecto, por así decirlo, independiente del tamaño de la muestra, en lugar de realizar una prueba que depende de él" (Bittner-Stephan, 2015, p.15). Así pues, el índice R dice algo *sobre lo alta que resulta ser la proporción de tasas de éxito de replicación típicas en el caso de la replicación exacta*, y lo hace en función de la potencia supuesta *operacionalizada* como la potencia media observada (mediana). De este modo, no sólo se puede examinar meta-analíticamente la replicabilidad de un área temática, sino también explorar la práctica de publicación de las revistas y el p-hacking potencialmente emergente. La investigación seria, según la autora, se caracteriza por una *pequeña diferencia* entre el índice R y la potencia real. Esto debe entenderse independientemente del hecho de que el valor absoluto de la potencia puede ser, no obstante, pequeño, por ejemplo si la potencia de los estudios es pequeña. El índice R adquiere así su significado *en un nivel relativo*. Una diferencia mayor entre el índice R y la potencia real indica que, como en el caso del p-hacking, se comunica demasiada potencia observada, es decir, que artificialmente hay demasiados valores *justo por debajo* del umbral de significación, lo que actualmente se aproxima mucho a una definición operativa de p-hacking. Esto va acompañado de una baja varianza de estos valores p (véase más adelante). Sin embargo, el índice R sólo cambia ligeramente en este caso, ya que la potencia observada inflada artificialmente se compensa más o menos con la creciente tasa de inflación durante el p-hacking (véase la ecuación 4.24). La potencia notificada sigue estando inflada. Lo que queda es una gran diferencia entre el índice R y la potencia notificada, que corresponde a una tasa de inflación creciente en el caso del p-hacking. En la práctica, la variación esperada de la muestra tiene un efecto de oscurecimiento, por lo que las fluctuaciones de valor son comunes y de esperar, especialmente si solo se dispone de unos pocos estudios para un metaanálisis. Schimmack (2017a) describe la clasificación del índice R según el siguiente esquema heurístico:

> „I consider an R-Index below 50 an F (fail). An R-Index in the 50s is aD, and an R-Index in the 60s is a C. An R-Index greater than 80 is considered an A."

Los conceptos de cálculo de *la potencia observada y la inflación* se presentan y discuten en Yuan y Maxwell (2005) y Schimmack (2015a), respectivamente. Se aplican las directrices adoptadas de Schimmack (véase la Tabla 4.5).

Tabla 4.5: *Índice R y potencia*

Índice *R*		**Condición**	**Comentario**
0%	(Mín.)	Mediana de "potencia observada" = 50% Tasa de éxito = 100%	No debería aparecer con datos empíricos, ya que la potencia de resultados significativos debería ser (Mín.) > 50%, es decir vemos aquí una variación de la potencia observada dependiente de la muestra con potencia obs. = 50%. También una variación muy límitada remite a sesgas.
22%	(Límite inferior)	H_0 es verdadera. Se informa solamente sobre resultados significativos ($p < 0.05$), es decir: Tasa de éxito = 100%. Mediana de potencia observada = 100%, tasa de inflación = 39%	La distribución de los valores *p* en caso de H_0 verdadera es uniforme. Índice *R* = 61 .. 39 Límite inferior realista con potencia de efecto de la población cerca de cero.
50%			Un índice R < 50 % implica que la potencia real es inferior al 50 %, lo que a su vez implica una potencia demasiado baja para la mayoría de las preguntas de investigación y, por tanto, cuestiona estos estudios.
100%	(Máx.)	Potencia = 100% Tasa de éxito = 100%	
Casos específicos			El índice *R* no es una medida de la potencia, pero muestra una relación monótona con la potencia, lo que permite clasificar la potencia de los estudios publicados al compararlos.
R > Potencia real		Potencia < 50%	Índice *R* conservador, con un índice *R* < 50 %, los estudios muestran muy poca potencia.
R < Potencia real		Potencia > 50%	

El límite inferior del 22 % (véase la Tabla 4.5) se calcula si la H_0 es cierta y solo se notifican los valores significativos. Éstos se encuentran en el rango de la tasa de error α. Si H_0 es cierta, los valores p siguen una distribución uniforme. El resto sigue las fórmulas anteriores según Schimmack (2016a). Expresado en R (`ptII_quan_classicstats_R-index_z-curve.r`) utilizamos valores arbitrarios dos veces:

```
# R Index arbitrary example R-Code
dats.1 <- data.frame(stat.test=rep("t",3),
    t       ype.test=rep("twoway",3),
            test.statistic=c(1.7,2,1.1),
            success=c(0,1,1),
            df.nominator=c(1,1,1),
            df.denominator=c(38,36,28),
            alpha=rep(0.05,3))
# R Index reproduce Schimmack spreadsheet data (single study)
dats.2 <- data.frame(stat.test="F",
            type.test="twoway",
            test.statistic=4.44,
            success=1,
            df.nominator=1,
            df.denominator=38,
            alpha=0.05)
```

Esto crea la salida:

```
> r.indx.dats.res <- R.index(dats=dats.1)
R-Index for a set of studies
Median (Observed Power) = 0.2792
Mean (Success Rate) = 0.6667
Inflation = 0.3875
R-Index = -0.1084
Side notes:
- Statistical Test = t
- alpha = 0.05
- effect = twoway
- no. of studies = 3
> r.indx.dats2.res <- R.index(dats=dats.2)
R-Index for a set of studies
Median (Observed Power) = 0.5303
Mean (Success Rate) = 1
Inflation = 0.4697
R-Index = 0.06052
Side notes:
- Statistical Test = F
- alpha = 0.05
- effect = twoway
- no. of studies = 1
NOTE: only single study, no set of studies
```

Este es otro ejemplo de Schimmack (2016a):

```
> # from Schimmack
> alpha <- 0.05
> # median uniform distribution 0.05 - 0
> # in case H0 is true -> median = 0.025
> median.p.unif <- punif(0.05)/2
> non.central.z <- qnorm(median.p.unif/2,lower.tail=FALSE)
> # = qnorm(1-median.p.unif/2)
> non.central.z
[1] 2.241403
> critical.z <- qnorm(1-alpha/2)
> critical.z
[1] 1.959964
> # because H0 = TRUE -> = type II error rate
> beta.errorrate <- pnorm(critical.z,non.central.z)
> beta.errorrate
[1] 0.389187
> median.obs.power = 1-beta.errorrate
> median.obs.power
[1] 0.610813
> inflation <- 1-median.obs.power
> R.index <- median.obs.power-inflation
> R.index
[1] 0.2216261
```

La mediana de la potencia observada es de 0.61 y el índice *R* es de 0.22 con un nivel convencional de 0.05. Schimmack y Brunner (2017b, con código *R* adjunto) amplían el debate sobre la replicación para incluir diferentes métodos (*curva p, curva z*, véase también Aplicaciones y consecuencias dentro de la comunidad científica, Schimmack, 2018c, 2018d, 2018b) para aproximar la potencia real de los estudios. Al hacerlo, el autor concluye que los estudios típicos de psicología suelen tener solo un 50% de potencia, con una gran heterogeneidad (es decir, variabilidad) entre los estudios. Esto significa que una cantidad considerable tiene muy poca potencia y, por lo tanto, puede producir *falsos negativos*, es decir, que no se encuentren los efectos existentes debido a la falta de potencia. Teniendo en cuenta el requisito del 80 % de potencia propuesto por Cohen (1969; 1992) aparece probable que la mayoría de los estudios, especialmente en psicología experimental (social), fracasen.

En resumen, el índice *R* no es una medida absoluta y exacta, pero es útil al *nivel relativo*. Depende de la calidad de los datos disponibles. Schimmack (2016b, 2017a) y Bittner-Stephan (2015, p.29.) informan en su trabajo de otros retos y limitaciones en el trabajo práctico con el índice *R*. Desde un punto de vista

epistemológico, parece problemático lo que la replicabilidad y un índice asociado significan realmente en términos concretos? ¿Se trata de si cabe esperar los mismos resultados en las mismas condiciones? Dado el hecho de la variación de la muestra con el tamaño de muestra típicamente pequeño de los estudios psicológicos, esto puede responderse con un claro "no lo sabemos". Precisamente cuando la potencia es pequeña, la variación de la muestra puede producir resultados sorprendentes. Por ejemplo, el estudio *50 sombras de gris* de Nosek, Spies y Motyl (2012) muestra cuánto pueden variar los resultados de los estudios. Por lo tanto, el uso del índice *R* solo tiene sentido cuando hay varios estudios. En la línea de los argumentos de Andrew Gelman, que se citan aquí varias veces en el libro, es importante buscar no solo los efectos, sino también la forma en que están condicionados. El índice *R* sigue muy centrado en la cuestión de los significados y la frecuencia de su aparición, sin especificarse más en términos de contenido. El nivel de contenido, la significatividad a nivel profesional y la direccionalidad de los efectos desempeñan un papel menor, si es que lo hacen, al igual que la complejidad de su interacción. Sin embargo, todos estos factores juntos afectan a la replicabilidad y a la precisión, es decir, al diseño y a la reducción de los errores de medición. De nuevo, no se trata de practicar un "o lo uno o lo otro", sino de integrar distintas fuentes de información para ver mejor el panorama general. El índice *R* no lo hace.

4.4.4.2 Curvas Z

El índice *R* no estima la replicabilidad per se; para ello, Brunner y Schimmack (2018a, 2018b) o Schimmack y Brunner (2017b, con código R) han desarrollado un instrumento llamado *curva Z*. Aquí, los valores *p* se transforman en valores *z* para integrar los resultados de diferentes estudios entre sí. El estadístico de prueba *Z* se reescala de varias maneras a un rango superior a un valor crítico c ≈ 1.96, es decir ≈ 2 desviaciones estándar para una prueba bilateral e inferior a 6, ya que a partir de $z > 6$ los valores se aproximan a cero y se vuelven difíciles de manejar (ibíd., p.17.). Estos valores *p* extremos se reintegran posteriormente. Debido a las variaciones del muestreo (tamaño de la muestra, tamaños del efecto), la distribución de *Z* consiste en una colección promediada de normales *truncadas*, en la que cada representante tiene un parámetro de *no centralidad mj* principalmente diferente. La distribución *Z* es, por tanto, un modelo mixto con *r* componentes (cada uno con parámetros de no centralidad *mj* y probabilidades asociadas *wj*) La distribución *Z* se distribuye normalmente, tiene la media *m* y la desviación estándar *1*. El reescalado garantiza que el área bajo la curva suma 1 y, por tanto, da como resultado áreas que pueden interpretarse como probabilidades. La función de densidad de probabilidad del estadístico de la prueba Z tiene, según Brunner y Schimmack (2018a, p.6, ecuación 6) tiene la fórmula

$$f(z) = \sum_{j=1}^{r} w_j \frac{\texttt{dnorm}(z - m_j)}{\texttt{pnorm}(6 - m_j) - \texttt{pnorm}(c - m_j)} \qquad \texttt{,para } c < z < 6 \tag{4.25}$$

`dnorm()` y `pnorm()` son las funciones R correspondientes para el cálculo de la densidad y la probabilidad acumulada, respectivamente. Así, por ejemplo, con el supuesto de una distribución normal estándar (μ = 0, σ = 1) son posibles los siguientes cálculos (`ptII_quan_classicstats_R--index_z-curve.r`):

```
# habitual alpha = 5% level
alpha <- 0.05
# density at modal value = mean = median
dnorm(0)
# density at typical critical significance level
dnorm(qnorm(1-alpha/2))
# cum prob at 1-alpha/2 (two sided test) quantile
pnorm(qnorm(1-alpha/2))
```

La función *f(z)* anterior se ajusta a lo largo de una función de estimación de densidad de núcleo implementada en R en `density()` para poder comparar diferentes distribuciones *Z* (por ejemplo, predicha,

observada, ...). Aquí, el *wj* y el *mj* se estiman para minimizar las desviaciones absolutas entre la estimación de la densidad del núcleo y la fórmula anterior. La probabilidad de rechazo (replicación) para $Z < 6$ se obtiene como el área bajo la curva por encima del umbral crítico basado en los pesos *wj* estimados y los parámetros de no centralidad *mj*. Mientras que los valores que entran en la ecuación 4.25 anterior están truncados, los valores para la probabilidad de rechazo no lo están, ya que según los autores (ibid., p.7)

> „The estimation of rejection probability upon replication for $Z < 6$ is the area under the curve above the critical value, with weights and non-centrality values from the curve tting step [...] Note that while the input data are censored both on the left and right as represented in Forumula 6, there is no truncation in Formula 7 because it represents the distribution of Z upon replication"

La ecuación es (ibíd., p.7, ecuación 7) para la potencia media de *Z* para el intervalo $Z < 6$

$$l = \sum_{j=1}^{r} \hat{w}_j(1 - \texttt{pnorm}(c - \hat{m}_j)) \quad \text{para Z} < 6 \tag{4.26}$$

Como paso final, los autores añaden los valores q extraídos anteriormente para $Z > 6$ en asumiendo que la probabilidad de un resultado estadísticamente significativo es cerca de 1.

$$\begin{aligned} Z &= (1-q)l + q \cdot 1 \\ &= q + (1-q)\sum_{j=1}^{r} \hat{w}_j(1 - \texttt{pnorm}(c - \hat{m}_j)) \end{aligned} \tag{4.27}$$

El valor *Z* indica entonces la probabilidad de rechazar la H_0 para una replicación exacta de una prueba seleccionada al azar, es decir, hace afirmaciones simplificadas sobre si un estudio parece replicable (= tasa de éxito). Brunner y Schimmack (2018a) investigan la aplicabilidad de este estadístico de prueba mediante simulación y también examinan los estudios de replicación del proyecto de replicación de la Open Science Collaboration (2015). Ahora se puede encontrar una implementación en R en el paquete de R `zcurve`. El ajuste se basa en la estimación de la densidad o en el algoritmo EM (Bartoš & Schimmack, 2020). Schimmack señala que, debido a las restricciones del índice *R*, debe combinarse con la prueba de varianza insuficiente TIVA (= test of insuffcient variance, 2014). El TIVA puede aplicarse a cualquier conjunto de estudios empíricos que realicen pruebas contra una H_0 e informen de los valores *p* evaluados para la significación. La TIVA cuantifica (ibíd., fórmula 1) la probabilidad de obtener resultados significativos en ausencia de varianza. Por un lado, se toma el cociente de la varianza observada frente a la varianza de la población (= 1) y el número k de valores z menos 1 (= df =grados de libertad). Por regla general, un conjunto de valores *z* insesgados tiene una varianza mínima de 1. Las varianzas de los valores *z* inferiores a 1 indican valores *z* distorsionados. Las razones para que falte una varianza son la falta de potencia, no informar de resultados no significativos y métodos de investigación cuestionables que sesgan los tamaños del efecto. Si k =número de valores z, entonces la TIVA se calcula según

$$\texttt{TIVA}_{\chi 2} = \texttt{varianza observada * (k .. 1), con } df \texttt{ = k - 1} \tag{4.27}$$

Schimmack (2014) ve el uso de TIVA de la siguiente manera

> „TIVA can be used to examine whether a set of published p-values was obtained with the help of questionable research practices. When p-values are converted into z-scores, the variance of z-scores should be greater or equal to 1. Insufficient variance suggests that questionable research practices were used to avoid publishing non-significant results; this includes simply not reporting failed studies."

Los datos del metaestudio OSC pueden reproducirse a través de la página de ayuda de `?zcurve`. Se trata exclusivamente de valores *z* de estudios originales. Como alternativa, pueden introducirse valores *p* que

pueden convertirse internamente en valores *z*. Ahora en primer lugar la forma general de convertir valores *p* en valores *z* - y viceversa (`ptII_quan_classicstats_R-index_z-curve.r`):

```
> # pvalue to z-score
> pval <- 0.06
> z.score <- qnorm(1-pval/2, mean=0, sd=1)
> pval
[1] 0.06
> z.score
[1] 1.880794
> # and back
> (1-pnorm(z.score, mean=0, sd=1))*2
[1] 0.06
```

Ahora al metaestudio OSC (según Schimmack, 2014, Ejemplo 2 o `?zcurve`) incl. el cálculo de de la TIVA (Los valores indicados por Schimmack (2014) son presumiblemente en parte erróneos o están incorrectamente impresos. Lamentablemente, en 2020 no se respondió a la correspondiente consulta al autor):

```
# Vohs et al. 2006
# http://www.sciencemag.org/content/314/5802/1154.short
pvals1 <- c(26,50,46,39,21,40,26,23,6)/1000
# no.5 = 299 (Schimmack, probably wrong!)
zzs1 <- c(223,196,199,206,230,206,223,228,273)/100
pvals1
zzs1
TIVA(z.scores=zzs1, type.test="twoway")
TIVA(pvals=pvals1, type.test="twoway")
# load existent data set
OSC.z
# EM with bootstrap
m.EM <- zcurve(OSC.z, method = "EM", bootstrap = 1000)
# KD2 with bootstrap
m.D <- zcurve(OSC.z, method = "density", bootstrap = 1000)
# output
summary(m.EM)
summary(m.D)
# plot the results
plot(m.EM)
plot(m.D)
```

La varianza de los nueve valores *z* es (véase también la salida de `TIVA()`)

```
> # variance z-scores
> var(zzs1)
[1] 0.05497778
```

y el estadístico χ^2 asociado (valores TIVA) da como resultado 0,4398 (df = 8) o p = 8,178e - 05. En resumen, se trata de un evento extremadamente raro y la varianza es mucho menor de lo que cabría esperar en un estudio no sesgado. Se puede ajustar y trazar una curva utilizando `zcurve()` (véase la Fig. 4.46). El gráfico superior muestra el cálculo con el algoritmo EM y 1000 repeticiones bootstrap; el gráfico inferior genera la versión KD2 según Bartoš y Schimmack (2020), también con 1000 repeticiones bootstrap.

Una aplicación del TIVA a los datos del estudio de Bem (2011a) demuestra la facilidad de aplicación de esta cuantificación. Bem (2011a) informa de que 9 de cada 10 estudios son estadísticamente significativos con valores p asociados. El siguiente código R calcula el TIVA a partir de esto utilizando la la función `TIVA()`. A esta función se le pueden pasar valores *z* o valores *p* a esta función. Si se pasan valores *p*, `TIVA()` los transforma de nuevo en valores *z* (Schimmack, 2014, Ejemplo 1, valores *p* de Bem, 2011a, p.421, Tab. 7).

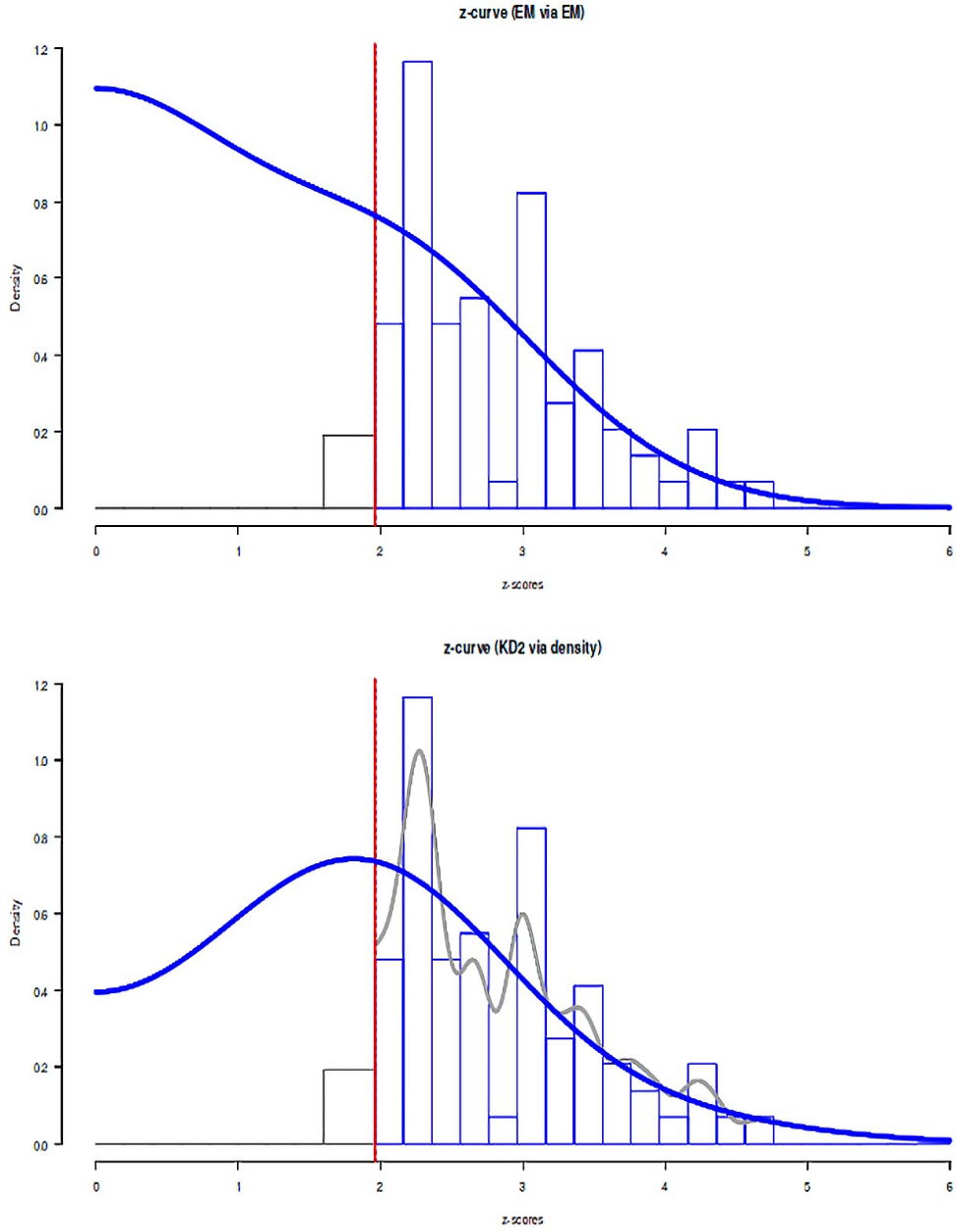


Figura 4.46. *Schimmack (2014), metaestudio OSC (curva z).*

El siguiente código R muestra esto (`ptII_quan_classicstats_R-index_z-curve.r`):

```
> # Bem study
> # 9 out of 10 studies significant at p < 0.05 one-tailed
> zzs <- c(2326,2336,2457,2197,2197,1787,1762,1305,1896,2878)/1000
> pvals <- c(10,9,7,14,14,37,39,96,29,2)/1000
> zzs
[1] 2.326 2.336 2.457 2.197 2.197 1.787 1.762 1.305 1.896 2.878
> pvals
[1] 0.010 0.009 0.007 0.014 0.014 0.037 0.039 0.096 0.029 0.002
> # chisquare value with k-1 df
> TIVA(z.scores=zzs, type.test="oneway")
TIVA Test for insufficient variance
Observed Variance = 0.1936
```

```
TIVA (Chi^2 statistic) = 1.742
p-value = 0.00508
df = 9
k = 10
NOTE: z-scores transformed to p-values

$res
TIVA p OV k
1.742189 0.005080087 0.1935765 10
$scores
pvals z.scores
1 0.010009275 2.326
2 0.009745621 2.336
3 0.007005134 2.457
4 0.014010223 2.197
5 0.014010223 2.197
6 0.036968744 1.787
7 0.039034649 1.762
8 0.095946424 1.305
9 0.028980022 1.896
10 0.002001026 2.878
> TIVA(pvals=pvals, type.test="oneway")
TIVA Test for insufficient variance
Observed Variance = 0.1953
TIVA (Chi^2 statistic) = 1.757
p-value = 0.005251
df = 9
k = 10
NOTE: p-values transformed to z-scores

$res
TIVA p OV k
1.757394 0.005250779 0.195266 10
$scores
pvals z.scores
1 0.010 2.326348
2 0.009 2.365618
3 0.007 2.457263
4 0.014 2.197286
5 0.014 2.197286
6 0.037 1.786613
7 0.039 1.762410
8 0.096 1.304685
9 0.029 1.895698
10 0.002 2.878162
```

La varianza de los valores z observados es $s^2 = 0.19$. La probabilidad del valor TIVA 1.757 (df = 9) es p = 0.0053. Esto significa que 1/0.0053 = 190.45, es decir 1 de 191 es de esperar para un conjunto aleatorio de los 10 estudios notificados (= valores *z*, véase más arriba), que se produzca una varianza de $s^2 = 0.19$. Basándose en este análisis, Schimmack (2014) llega a la conclusión,

> „This outcome cannot be attributed to publication bias because all studies were published in a single article. Thus, TIVA supports the hypothesis that the insufficient variance in Bem's z-scores is the result of questionable research methods and that the reported effect size of d = .2 is inflated. The presence of bias does not imply that the true effect size is 0, but it does strongly suggest that the true effect size is smaller than the average effect size in a set of studies with insuffcient variance."

Así pues, los ámbitos de la *potencia* y las *réplicas* resultan muy significativos en la práctica de la investigación y no se limitan a la simple exigencia de una mayor potencia y resultados replicables. Sin embargo, debemos recordar que la investigación siempre puede equivocarse, tanto en los estudios como en las réplicas y los metaanálisis. Parece sensato que los avances en este campo produzcan medidas de cuantificación para poder evaluar mejor los estudios y la práctica de publicación de las revistas de forma retrospectiva y metaanalítica. Sería bueno añadir a estas medidas resultados adicionales para el mejor caso

frente al peor caso, así como datos de distribución, ya que las estimaciones puntuales generalmente sólo proporcionan una imagen inadecuada de una situación.

En resumen, existen herramientas estadísticas para identificar aproximadamente los sesgos típicos. La tendencia a retener los resultados supuestamente positivos y a ignorar las falsificaciones potenciales surgen aquí como sucesos frecuentes. Esto se apoya en una práctica de publicación indiscutible de muchas revistas, que sólo publican resultados positivos y ninguna réplica.

4.4.5 (Auto-) Engaños

Entre los (auto)engaños se encuentra el fenómeno de la falacia de proyección mental descrito por Jaynes (1988a), que utiliza en el contexto de una crítica a la interpretación de Copenhague de la física cuántica (Jaynes, 1996-08-07). La falacia cognitivo-emocional descrita tiene lugar cuando alguien supone que su propia percepción del mundo se corresponde exactamente con él. La alegoría socrática de la caverna (Platón, 1994) no es otra cosa que esto cuando se interpretan las sombras en la pared creadas por la luz, pero nunca se experimenta la propia luz.

El mismo problema se da básicamente en la interpretación de los datos, por lo que, en consecuencia, el análisis secuencial (véase también el capítulo 11.9.6 sobre los errores típicos al llevarlo a cabo) es mejor realizarlo en grupo para escapar, al menos parcialmente, de la propia ceguera operativa. Sin embargo, puede ocurrir que este acto engañoso no se limite a una persona, sino que se extienda a varias. Un ejemplo bien conocido de la psicología social apunta a los grupos mínimos (Tajfel, 1970), los desplazamientos de la percepción social hacia la extremista de los grupos (groupthink) según Janis (1971) o el fenómeno del desplazamiento arriesgado según Stoner (1961) en torno al presidente Kennedy en los años sesenta. Este último por poco condujo a una guerra nuclear. Las características típicas son la retención de creencias a pesar de los hechos evidentes en contra y la tendencia a tratar de confirmar las propias suposiciones en lugar de cuestionarlas críticamente. La era de las noticias falsas nos recuerda a diario esta peligrosa forma de formación de opinión subjetiva y sus efectos en la política, la economía, la sociedad, etc. La psicología social enumera muchos otros efectos perceptivos similares en contextos sociales que pueden tener efectos distorsionadores. Los componentes centrales de todos estos efectos son, en su mayoría, el procesamiento defectuoso de la información y las conclusiones basadas en él, creencias no basadas en la experiencia real, errores de atribución, profecías autocumplidas, estereotipos y la sobreestimación o subestimación sistemática de los efectos.

4.4.6 Estimación insesgada de las varianzas muestrales

En los cursos introductorios de estadística clásica, ¿quién no se ha preguntado alguna vez por la *cuestión de las estimaciones insesgadas o la precisión esperada de las varianzas*? La cuestión de la precisión, o mejor la cuestión del *estimador insesgado* („unbiased estimator“), se considera uno de los santos griales de la estadística clásica – junto con *la consistencia, la suficiencia* y *la eficiencia (asintótica)* – para evaluar la calidad de los parámetros estimados. Del físico E.T. Jaynes, representante de la estadística objetiva de Bayes, que trabaja con pruebas de realidad crítica, proceden contra-argumentos sustanciales a los procedimientos mencionados en todos los libros de texto para corregir la varianza con el fin de obtener estimaciones insesgadas en condiciones finitas (Jaynes, 2003, cap. 17). Los contra-argumentos proceden de la lógica y la teoría de la información, aunque pueden llevarse a cabo matemáticamente. A Jaynes no le preocupa la aritmética, sino comprender qué es realmente una estimación errónea o sesgada, si debe o no llamarse error en el sentido de estar equivocado y cuáles son sus consecuencias – y qué corrección surge de la comprensión respectiva.

La forma clásica resulta de la siguiente manera (Jaynes, 2003; Oliphant, 2006-12-05, p.1.) – *si se conoce el valor esperado de la varianza*, entonces para la varianza muestral *Var[X]* de una variable aleatoria *X* distribuida normal resulta la estimación Maximum-Likelihood:

$$\mathbb{E}\mathrm{Var}(X)] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(X_i - \mu)^2] \tag{4.29}$$

y para el valor esparado fijo

$$\mu = \mathbb{E}[X_i] \tag{4.30}$$

Suponiendo que *la variable X* es distribuida normal, sigue que μ^ y σ^ son independientes, (n -1)*(σ^²/σ^) es distribudo de modo χ^2 con df = n-1 grados de libertad. Se calcula la desviación estándar convencionalmente como raiz cuadrada de σ^² sin justificarlo (salvo por convención) por la estimación no-sesgada (véase más abajo; Oliphant, 2006-12-05). Se puede deducir de estes datos por ejemplo los intervalos de confianza.

En caso de que *el valor estimada de la varianza sea desconocido a causa de la media incógnita* μ_0 se tiene que estimar la media basado en la muestra:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}(X_i) \tag{4.31}$$

El valor esperado de la muestra estimada y su varianza se encuentra como

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}[X] \tag{4.32}$$

$$\mathrm{Var}[\hat{\mu}] = \frac{1}{n}\mathrm{Var}[X] \tag{4.33}$$

La pregunta que Jaynes, Oliphant y otros se plantean aquí es qué corrección o qué procedimiento de estimación aprovecha mejor la información de los datos y la estima con mayor eficacia, en función del problema específico. La estimación UB no es esto.

La tabla 4.6 enumera tres métodos de estimación diferentes según Oliphant (2006-12-05, p.3) para la estimación de la varianza, su valor esperado y el error cuadrático medio (mean square error; MSE) asociado:

- máxima verosimilitud (Maximum-Likelihood; ML)
- estimación insesgada (unbiased estimation; UB)
- error medio cuadrático mínimo (Minimal Mean-Square Error; MMSE).

Tabla 4.6. *Métodos de estimar la varianza*

Methode	$\hat{\sigma}^2$	$\mathbb{E}[\hat{\sigma}^2]$	$MSE[\hat{\sigma}^2]$
ML	$\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$	$\frac{n-1}{n}\sigma^2$	$\frac{2n-1}{n^2}\sigma^4$
UB	$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu}^2)$	σ^2	$\frac{2}{n-1}\sigma^4$
MMSE	$\frac{1}{n+1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$	$\frac{n-1}{n+1}\sigma^2$	$\frac{2}{n+1}\sigma^4$

Para la estimación insesgada de la varianza, resulta la conocida fórmula corregida por n-1 en el denominador

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2 \tag{4.34}$$

De este modo se garantiza lo siguiente:

$$\mathbb{E}[\hat{\sigma}^2] = \hat{\sigma}^2 \tag{4.35}$$

El error cuadrático medio (MSE) tiene la siguiente identidad si E[θ^] corresponde al sesgo (bias) y X_i procede de una distribución normal con media μ y varianza σ^2:

$$\begin{aligned} MSE[\hat{\theta}] &\equiv \mathbb{E}[(\hat{\theta} - \theta)^2] \\ MSE[\hat{\theta}] &= \mathrm{Var}[\hat{\theta}] + bias^2 \\ &= \mathrm{Var}[\hat{\theta}] + (\mathbb{E}[\hat{\theta} - \theta])^2 \\ \mathbb{E}[(\hat{\theta} - \theta)^2] &= \mathbb{E}[\hat{\theta}^2] - 2\theta\mathbb{E}[\hat{\theta}] + \theta^2 \end{aligned} \tag{4.36}$$

De la última expresión se deduce que cuando $E[(\hat{\theta} - \theta)^2]$ se aproxima a cero, la varianza y el MSE convergen y se hacen idénticos en cero. Para la desviación estándar, la tabla 4.7 enumera las fórmulas complementarias de la varianza. Un código R reproduce los gráficos de Oliphant (2006-12-05) para la varianza y la desviación estándar (`ptII_quan_classicstats_varianceestimation.r`) y el resultado de la Figura 4.47:

```
>   # if sigma is known
>   varest(x=rnorm(n=50, mean=100, sd=10), sigma=10, LOG=FALSE)
label       n   shat2       s2.normmean E.shat2     MSE.shat2 s2.normRMSE
ML ML       50  90.57460    0.9800000    98.00000   396.0000  0.1989975
UB UB       50  92.42306    1.0000000   100.00000   408.1633  0.2020305
MMSE MMSE   50  88.79862    0.9607843    96.07843   392.1569  0.1980295
      shat      s.normmean E.shat     MSE.shat  s.normRMSE
ML    9.517069  0.9849119   9.849119 1.017612  0.1008767
UB    9.662862  1.0000000  10.000000 1.025561  0.1012700
MMSE 9.564770   0.9949113   9.949113 1.015150  0.1007546

> # if sigma is unknown
> varest(x=rnorm(n=50, mean=100, sd=10), sigma=NA, LOG=FALSE)
label      n  shat2    shat
ML ML      50 79.20965 8.899980
UB UB      50 80.82617 9.036321
MMSE MMSE 50 77.65652 8.944589
> plot.varest(plotwhat="s2")
> plot.varest(plotwhat="s")
```

Tabla 4.7. *Métodos de estimar la desviación estándar*

Methode	$\hat{\sigma}$	$E[\hat{\sigma}]$	$MSE[\hat{\sigma}]$
ML	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2}$	$t_n\sqrt{\frac{2}{n}}\sigma$	$2\sigma^2\left(1 - t_n\sqrt{\frac{2}{n}} - \frac{1}{2n}\right)$
UB	$\frac{1}{t_n}\sqrt{\frac{1}{2}\sum_{i=1}^{n}(X_i - \hat{\mu}^2)}$	σ	$\sigma^2\left(\frac{n-1}{2t_n^2} - 1\right)$
MMSE	$\frac{t_n}{n-1}\sqrt{2\sum_{i=1}^{n}(X_i - \hat{\mu})^2}$	$t_n\sqrt{\frac{2}{n-1}}\sigma$	$\sigma^2\left(1 - \frac{2t_n^2}{n-1}\right)$

Con *muestras infinitamente grandes* – o con conocimiento de los valores poblacionales relevantes (= valor medio) – las correcciones introducidas son insignificantes. Tanto las variantes no corregidas como las corregidas son asintóticamente tolerantes a las expectativas, ya que para ambas el valor límite para $n \to \infty$ se encoge hacia cero, es decir, la varianza se determina con exactitud y precisión infinitas.

Con *muestras pequeñas*, sin embargo, esta corrección *no es trivial*, ya que se incluye en el cálculo de los grados de libertad, que a su vez determinan los valores *p*, los límites de confianza, etc. Aquí es donde entra Jaynes (2003), que no critica la necesidad de una corrección ni la fidelidad asintótica a las expectativas,

sino la forma en que se aplica la corrección para muestras finitas concretas. Jaynes subraya que el estimador UB es *ineficiente* y no se ajusta necesariamente a una muestra concreta, ni siquiera proporciona el mejor estimador para una muestra concreta. Por el contrario, Jaynes (2003, 511.) deriva como ejemplo el estimador MMSE (véase la Tab. 4.6), que requiere sólo la mitad del tamaño de la muestra para la misma precisión que el estimador UB clásico. En sus explicaciones, Jaynes adopta una perspectiva teórico-informacional y no entiende los *errores* como "equivocaciones" que hay que eliminar a toda costa sino que examina la información contenida en ellos que da motivos para minimizar los errores total. „If it had been called instead the 'component of error orthogonal to the variance', as suggested by the Pythagorean form of ([equation] 17.2), it would have been clear to all that these two contributions to the error are on an equal footing; it is folly to decrease one at the expense of increasing the other" ("Si se hubiera llamado en cambio "componente del error ortogonal a la varianza", como sugiere la forma pitagórica de ([ecuación] 17.2), todos habríamos tenido claro que estas dos contribuciones al error se encuentran en el mismo nivel; es una locura disminuir una a expensas de aumentar la otra" (Jaynes, 2003, p.514). Jaynes no apunta a condiciones de infinito poco realistas, con las que la estadística clásica prefiere trabajar, aunque, como se ha mencionado, cada uno de los métodos de estimación presentados es verdadero desde el punto de vista de las expectativas y cualquier corrección se vuelve insignificante para tamaños de muestra muy grandes. Su crítica no es matemática, sino lógica o conceptual. Así, según Jaynes (ibíd.), el error total no se ve afectado por la corrección 1/(n - 1), sino de 1/ n + 1. En el caso de la regresión, la corrección da como resultado 1/(n - p + 2) en lugar de 1/(n - p -1) (ibíd.). Además, señala que su sugerencia es sólo una de muchas y que hay otras, dependiendo de la situación de los datos. Por tanto, la vía clásica es matemáticamente correcta, pero teóricamente ineficaz desde el punto de vista de la información y, por tanto, impracticable. Jaynes no limita el punto de vista a una solución general válida en todas las circunstancias, sino que se ocupa de la mejor solución para un problema claramente definido dada una muestra finita definida (ibid., p.515, cursiva en el original):

> "Furthermore, asymptotic behavior of an estimator is not really relevant, because the real problem is always to do the best we can with a nite data set, therefore the important question is not *whether* an estimator tends to be the true value, but *how rapidly* it does so."

Las causas subyacentes de este error de razonamiento residen de nuevo en las convenciones, ya que – como dice Oliphant (2006-12-05, p.2) – „[w]e are supposed to believe that this is preferable to an estimator that instead minimizes some other metric such as the mean-squared error (which includes both bias and variance)." ("Se supone que debemos creer que esto es preferible a un estimador que minimice alguna otra métrica, como el error cuadrático medio (que incluye tanto el sesgo como la varianza)". Sin embargo, el paquete `simr` de R permite muchas más posibilidades que las que se muestran aquí. Con `?modify`, por ejemplo, se puede cambiar cualquier área de los objetos `lmer` para adaptarlos a los análisis de potencia. Con `?tests`, se dispone de diversas variantes de pruebas (incluido el bootstrap paramétrico); y los propios datos pueden modificarse con `?getData`. Uno de estos estimadores alternativos a σ^{2ML} o σ^{2UB} es el estimador de minimal mean-square error (MMSE). Para $n > 1$, puede demostrarse que tanto el estimador MMSE como el ML dan mejores resultados que el estimador UB, ya que minimizan más el error total:

$$\mathrm{MSE}[\hat{\sigma}^2_{\mathrm{MMSE}}] < \mathrm{MSE}[\hat{\sigma}^2_{\mathrm{ML}}] < \mathrm{MSE}[\hat{\sigma}^2_{\mathrm{UB}}] \qquad (4.37)$$

Esto significa que con el criterio MSE o ML, los estimadores sesgados proporcionan mejores estimaciones que la estimación UB ineficiente. Oliphant (2006-12-05, p.4) muestra las curvas para esto y da más detalles para las estimaciones de las desviaciones estándar en su artículo así como para las estimaciones correspondientes según el teorema de Bayes. Con respecto a este último, cabe señalar que aquí se presentan otros estimadores, ya que no sólo minimizan una integral que recorre los datos como los estimadores no bayesianos (véase la Tab.??), sino también una integral que abarca el conocimiento previo. Esto significa que hay dos problemas de optimización diferentes y, en última instancia, dos tipos de enunciados, aunque no se puede suponer que lleguen a las mismas conclusiones.

Por parte de la estadística clásica, estos importantes argumentos de Jaynes no se han retomado hasta ahora ni se ha reflexionado nunca sobre ellos. Nosotros, los autores, al menos no conocemos ningún artículo

de estadística clásica o incluso libro de texto de estadística clásica o psicología que aborde el tema o incluso represente el punto de vista de Jaynes. Sin embargo, el tema en sí no es desconocido, como muestra el breve artículo de Diepgen (1999) para la escuela superior, que, sin embargo, proviene del contexto de una revista escolar y no de la universidad (¡!). En general, el tema no se discute en absoluto en la literatura o los libros de texto y más bien se puede encontrar en las entradas del blog (2017).

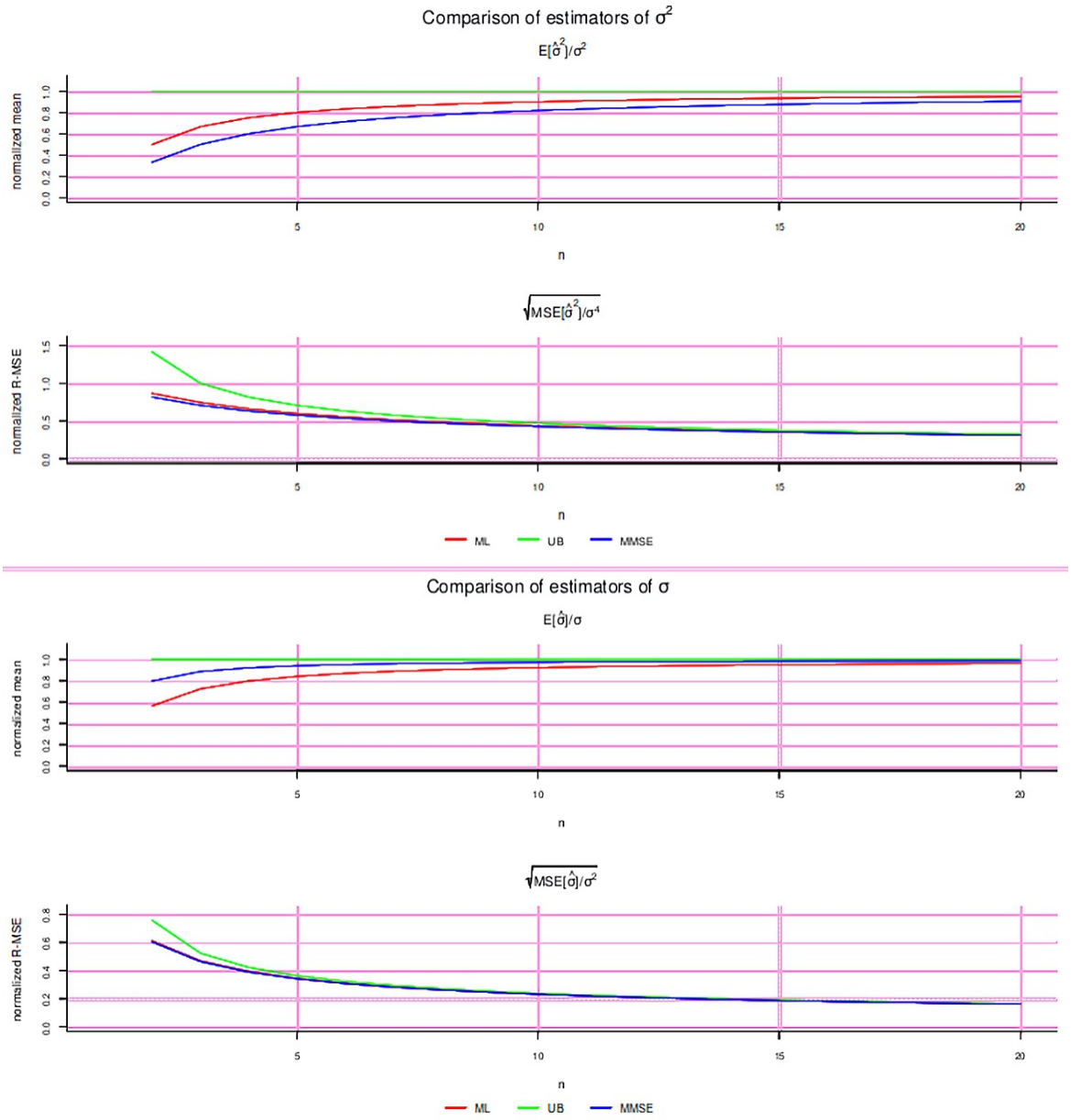


Figura 4.47. *Comparación de varias estimaciones de varianza (Oliphant, 2006-12-05)*

Una consulta a Van Horn (2004), que al fin y al cabo publica la fe de erratas del libro de Jaynes, dio como resultado la respuesta (comunicación personal de los autores con Van Horn) de que Jaynes publicó sus argumentos lógica y matemáticamente completamente sólidos. Jaynes (2003, p.515.) afirma además que el criterio de error no es invariante con respecto a ningún cambio de parámetros. Así, puede ocurrir que "el cuadrado de una estimación insesgada de α no sea una estimación insesgada de α^2". A medida que aumenta α^k, las diferencias pueden llegar a ser desproporcionadamente grandes, lo que lleva a conclusiones incoherentes con datos fijos a pesar de ser idénticos, ya que los problemas específicos no determinan la elección de k.

Hay que tener en cuenta que no se deben aceptar todos los supuestos sin cuestionarlos, sobre todo cuando se trata de cuestiones centrales y éstas se encuentran de forma irreflexiva en libros de texto muy conocidos. Es mucho más esclarecedor pensar fuera de la caja y aprender de otras profesiones para poder identificar algunos problemas. Existen argumentos similares sobre la Maximum Likelihood (Minka, 1998), a menudo desde una perspectiva bayesiana y un razonamiento coherente. En este punto, Minka (1998, p.4) hace hincapié en la necesidad de encontrar soluciones relacionadas con el problema concreto y no realizar análisis generales en el supuesto erróneo de que éstos son validos siempre y en toda circunstancia:

> „This provides a reminder that neither heuristic can be recommended universally. More generally, [it] is not guaranteed that even one of the heuristics will be sensible for your problem."

4.4.7 Aleatorización

Eid, Gollwitzer y Schmitt (2010, p.59) se refieren a la aleatorización como la "reina de las técnicas de control", ya que excluye la influencia de todas las variables de confusión imaginables ligadas a la persona. Hays (1974, p. 73) define la aleatorización de la siguiente manera:

> „A method of drawing samples such that each and every distinct sample of the same size N has exactly the same probability of being selected is called simple random sampling."

o en una formulación inversa (ibíd.)

> „Simple random sampling is a process of selecting elementary events for observation in such a way that each and every elementary event has precisely the same probability of being included in any sample of N observations."

Añade que las probabilidades respectivas pueden ajustarse al contexto, de modo que, por ejemplo, si es necesario, se produzca una determinada relación entre las probabilidades de ocurrencia. Asimismo, se pueden introducir dependencias entre los "sucesos generados aleatoriamente". La aleatorización suele justificarse por la ley de los grandes números. En el excursus del capítulo 4.3.5 sobre el tema de la simulación, ya se examinó la aleatorización – sacar con o sin retroceso (bootstrap, permutación) – a nivel práctico, por ejemplo, cuando se trata de sondear límites de confianza de estimaciones de parámetros o de estimar progresiones. Básicamente, la aleatorización intenta establecer algo que nunca se corresponde con la realidad situacional concreta, a saber, introducir una distribución equitativa de las variables influyentes. Esto es especialmente evidente en el experimento clásico, que se diferencia del cuasi-experimento en que la muestra se genera de forma totalmente aleatoria, lo que significa que, independientemente de la razón situacional, cada persona tiene la misma probabilidad ser asignada a una de las diversas condiciones experimentales. Mientras que, por un lado, se ejerce el máximo control experimental, por otro, esto lleva prácticamente a desechar la información existente en favor de la aleatorización. Si bien esto puede tener un efecto positivo en un experimento de laboratorio estrictamente definido en el marco del estudio, la validez ecológica y, por tanto, la transferibilidad a la vida normal se resienten. Jaynes (2003, p. 72) ofrece un análisis muy acertado de esta situación problemática en una digresión sobre la aleatorización en el contexto de las técnicas de muestreo. En primer lugar, Jaynes redefine el concepto de aleatorización y lo precisa (ibid., p.73, cursiva en el original):

> „This term is, evidently. a euphemism, whose real meaning is: *deliberately throwing away relevant information when it becomes too complicated for us to handle.*"

Jaynes describe lo que quiere decir con esto utilizando el ejemplo de una urna y sacar bolas con retroceso (volver a ponerlas un su sitio). Es prácticamente imposible calcular con exactitud la probabilidad de sacar cualquier bola de una urna, aunque se dispusiera de todos los datos y de toda la potencia informática del mundo. Jaynes no cree que esto sea teóricamente imposible, pero es demasiado complicado.

El proceso habitual de aleatorización, en este caso concreto, consiste en agitar la urna de tal manera que el complejo problema se simplifica en un experimento clásico de Bernoulli. También se podría decir que el sistema pasa a un estado de alta entropía. Sin embargo, no está claro cuánto tiempo hay que agitar la urna para alcanzar este estado. Una vez más, falta el criterio absoluto, de modo que ninguna bola está en principio en desventaja para cumplir el requisito de la aleatorización, es decir, la igualdad de probabilidades iniciales.A sus afirmaciones deliberadamente polémicas, Jaynes añade otras (ibid.):

> „For some, declaring a problem to be 'randomized' is an incantation with the same purpose and effect as those uttered by an exorcist to drive out evil spirits; i.e. it cleanses their subsequent calculations and renders them immune to criticism. We agnostics often envy the True Believer, who thus acquires so easily that sense of security which is forever denied to us."

Jaynes admite, sin embargo, que este procedimiento conduce a menudo a aproximaciones útiles a la solución correcta. Es decir, las complejas y relevantes influencias individuales sólo tienen un pequeño efecto numérico sobre la solución siempre que n sea suficientemente pequeño (ibid., p.75), porque

> „for small n, this approximation will be quite good; but for large n these small errors can accumulate (depending on exactly how we shake the urn, etc.) to the point where [a formula] is misleading."

Se introduce una cierta incertidumbre, difícilmente especificable, que se acumula en errores y distorsiones fuera de control, porque el mecanismo subyacente de estos errores es desconocido. Por tanto, existe un problema metodológico fundamental y no hay ninguna prueba en sentido estricto de que este enfoque sea el correcto, salvo para decir que en el caso de la urna anterior, la duración de la agitación conduce a una mejor aproximación a una distribución igual de las probabilidades. Por cierto la agitación no produce azar, ya que este término carece completamente de sentido en el mundo real. No hay ningún principio aplicable detrás de él y tampoco está claramente denotado. Para Jaynes, el término cae en el terreno de las falacias de proyección mental (Jaynes, 1988a), es decir, autoengaños (ibíd., cursiva en el original):

> „What shaking accomplishes is very different. It does not affect *Nature's* workings in any way; it only ensures that no *human* is able to exert any willful influence on the result. Therefor nobody can be charged with 'fixing' the outcome."

Esto se corresponde con la definición de supuesta objetividad al introducir un proceso que influye en los datos, pero no de tal manera que la persona misma interactúe directamente con el objeto, sino sólo los procesos creados por el hombre, que, de nuevo, nunca parecen cuestionar la objetividad per se. Una vez más, Jaynes se preocupa por los principios y no por la pragmática cotidiana. No se critica el procedimiento en sí, ya que en las matemáticas aplicadas y en contextos reales siempre hay que hacer concesiones para llegar a aproximaciones razonables a las soluciones. Esto también se aplica al campo de la teoría de la probabilidad. Sin embargo, la creencia de que este procedimiento es algo más que una aproximación a algo que básicamente no se entiende tiene un efecto fatal. Las ecuaciones no se hacen más exactas mediante la aleatorización y tampoco se entiende mejor el tema, ya que la argumentación no tiene lugar a nivel de contenido y los datos generados son simulados y no de naturaleza real. Esto prohíbe los supuestos que contengan la palabra "prueba" o "evidencia", ya que no hay nada que probar con la aleatorización, especialmente no "la identidad de la probabilidad y la frecuencia límite" (ibíd., p. 75). Los teoremas límite, como la ley de los grandes números, no deben confundirse con verdades sobre la realidad (ibíd., p. 75).

> „Our point is that these theorems are valid properties *of the abstract mathematical model that was defined and analyzed*. The issue is: to what extent does that model resemble the real world? It is probably safe to say that no limit theorem is directly applicable in the real world, simply because no mathematical model captures every circumstance that is relevant in the real world. Anyone who believes that he is proving things about the real world, is a victim of the mind projection fallacy".

Tabla 4.8: *Diseño en bloques aleatorizados*

	Cantidad		
	Bloque	Factores	Vueltas
Grupos no homogéneos	1	k	$L_1 \cdot L_2 \cdot \ldots \cdot L_k$
	2	k	$L_1 \cdot L_2 \cdot \ldots \cdot L_k$
	3	k	$L_1 \cdot L_2 \cdot \ldots \cdot L_k$
	.....		$L_1 \cdot L_2 \cdot \ldots \cdot L_k$
	n	k	$L_1 \cdot L_2 \cdot \ldots \cdot L_k$

L_k = Cantidad de condiciones (niveles) por factor

¿Qué significa esto para la práctica de la investigación? Como ya se ha dicho en otros temas (por ejemplo, las estimaciones insesgadas, véase el capítulo 4.4.6), las matemáticas como tales no son criticables. Más bien hay que cuestionar críticamente su interpretación y, en consecuencia, las conclusiones relacionadas con su contenido. O, dicho de otro modo, se puede y se debe recurrir a la aleatorización si realmente no hay otra forma de obtener resultados. Pero creer entonces que se han llevado a cabo réplicas de estudios o incluso que se han obtenido resultados superiores a ellos o que incluso representan la realidad de forma más correcta que los datos reales sería un caso de flagrante autoengaño.

Los ejemplos clásicos anteriores, que suelen girar en torno a sacar bolas de urnas o lanzar una moneda al aire, no se dan, por supuesto, en la práctica cotidiana de la investigación. Pero constituyen una buena metáfora para comprender el proceso básico de cómo pueden producirse los errores. A veces incluso reflejan muy bien las condiciones de la vida real, por ejemplo, en física al contar la desintegración radiactiva, la garantía de calidad en la industria o las encuestas, de modo que "uno está literalmente extrayendo de una población real finita" (ibid, p.82). Sin embargo, otros contextos reflejan proporciones tan diferentes que la comparación con urnas y monedas. Por eso resulta bastante peligrosa cuando se utiliza estes modelos. Las interacciones sociales, por ejemplo, se encuentran entre ellos, al igual que todo el campo de los trastornos mentales y probablemente la mayoría de los problemas de la investigación educativa empírica. Jaynes (ibid., p.531) demuestra la crítica a nivel numérico con un ejemplo concreto. Por ejemplo, en R se pueden encontrar los paquetes `randomizeR` y `randomizr` para llevar a cabo planes especiales de aleatorización según determinados criterios, aparte de la simple aleatorización con `sample()`.

Se trata, por ejemplo, de requisitos de los ensayos clínicos como la agrupación o la evitación del sesgo de selección, que puede dar lugar a una asignación no aleatoria de las personas a las condiciones de tratamiento y falsear así los resultados del estudio. En el extremo, el sesgo de selección corresponde a la falta de validez externa, es decir, una discrepancia entre la muestra del estudio y la población clínica real a la que se aplican las conclusiones extraídas de los datos empíricos. Kahan, Rehal y Cro (2015) discuten este tema, muy relevante para los ensayos clínicos. `randomizeR` y `randomizr` se explican en varias viñetas (Coppock, Cooper & Fultz, 2019-04-23; Schindler, Uschner, Msnolov, Pham, Hilgers & Hilfers, 2018-06-15) y artículos (Uschner, Schindler, Heussen & Hilfers, 2018). `randomizeR` también permite una comparación directa y gráfica entre diferentes procedimientos de aleatorización. Vamos a entrar en algunas de las muy amplias posibilidades en extractos. En realidad, R viene con todo lo necesario para aleatorizar. Esto incluye el generador interno de aleatorización, la función `sample()` así como las diversas funciones de distribución.

Todas las cuales ofrecen la posibilidad de extraer números aleatorios de la distribución respectiva. En el sitio web del Center for Biostatistics de la Icahn School ofMedicine at Mount Sinai (ISMMS), hay una aplicación web de R en línea que implementa la aleatorización robusta para ensayos clínicos (Clinical Research APPS, 2017; Tu & Benn, 2017) y que se programó con el paquete R `shiny`, que permite programar aplicaciones web interactivas a partir de código R. La aleatorización como tal puede caracterizarse con algunos puntos clave a la hora de cuando se trata de sacar números aleatorios:

- *Extracción con retroceso* (= bootstrap)
- *Extracción sin retroceso* (= permutación)
- *Extracción sobre la base de probabilidades predefinidas* o sobre la base de una especificación o suposición de distribución específica o suposición
- *Extracción basada en requisitos específicos y condiciones marco no aleatorias*, por ejemplo en experimentos biológicos o en ciencias agrarias.

En el caso de los diseños de bloques aleatorizados (= RBD; randomized block design) por ejemplo, las combinaciones de niveles de factores (=covariables) se agrupan en bloques y estos diseños de bloques se repiten varias veces. Dentro de los bloques tiene lugar una asignación aleatoria, por ejemplo, a qué sección de un terreno se asigna qué combinación (=bloque) o qué combinación de niveles de factores es eficaz. En un ejemplo ficticio, se investiga cómo afectan al crecimiento de las plantas la cantidad de agua con seis condiciones diferentes y un determinado fertilizante biológico en tres concentraciones distintas. En este caso, se crean 6*3 = 18 condiciones de estudio y cada condición de estudio bloqueada se repite tres veces, es decir, 18*3 = 54 condiciones de estudio en total. En la Tabla 4.8 se presenta el diseño de bloques aleatorizados de forma general. Los bloques en el diseño de bloques aleatorios no tienen un significado más profundo y se consideran factores aleatorios y no son explícitamente condiciones de tratamiento. Pueden ser, por ejemplo, la ubicación, la planta, el lote, el tiempo, la persona que realiza el trabajo, etc. Se presupone una correlación cero entre el bloque y las condiciones. La asignación de las unidades de estudio a las condiciones dentro de los bloques es aleatoria. Cada condición debe darse al menos una vez por repetición. Es posible cualquier número de réplicas, lo que puede dar lugar a que determinadas condiciones de tratamiento se den con más frecuencia que otras. Sin embargo, cada condición tiene la misma probabilidad de ser utilizada. El objetivo de todo este esfuerzo es garantizar que las diferencias entre condiciones puedan atribuirse de la forma más causal posible sólo a estas condiciones y que el esfuerzo siga siendo bajo. Básicamente, se trata de experimentos independientes completamente aleatorizados por bloque. El diseño básico puede modificarse y complicarse como se desee. Una realización es posible con `randomizeR`. La tabla 4.9 visualiza adicionalmente el esquema general de cuadrados latinos, un diseño de bloques con dos variables de bloque ortogonales, aquí: Grupo y Tiempo (=repetición de la medición). El ejemplo declina permutativamente a través de todas las secuencias de condiciones. Cada condición se da en todas las filas y columnas. Este diseño se utiliza en experimentos agrícolas, entre otros, y rara vez en la investigación en ciencias sociales. Encontrará explicaciones más detalladas y casos especiales de los cuadrados latinos en Bortz (1993, capítulo 11.2). Aquí es interesante la relación entre los cuadrados latinos y los diseños experimentales completos (ibíd., p.366, Tab. 11.19). En comparación con el diseño experimental completo, el cuadrado latino requiere menos personas debido a la permutación y, por tanto, difiere de él en un factor p, si p es el número de condiciones por bloque. En cuanto a los efectos principales, existe un equilibrio completo, mientras las interacciones sólo están parcialmente equilibradas.

Se empieza con la definición de una muestra y el valor inicial para el generador aleatorio (`ptII_quan_classicstats_randomization.r`).

```
# sample
nsamples <- 2
npersample <- 30
stotal <- nsamples*npersample
seed <- 07987
set.seed(seed)
```

Tabla 4.9: *Diseño "cuadrado latino"*

Grupo*	Tiempo			
	t_1	t_2	t_3	t_x
1	b_1	b_2	b_3	...
2	b_2	b_1	b_3	...
3	b_3	b_2	b_1	...
...	...	...	...	...
k	...	...	...	...

* Aleatorización completa de niveles de condiciones b_j por grupo y los puntos de tiempo t_x (= tratamiento)

En primer lugar, nos ocuparemos de las funciones estándar de R antes de pasar a los paquetes específicos. Existen diferentes formas de generar aleatoriedad. Por ejemplo, la aleatoriedad puede realizarse utilizando números aleatorios y se puede realizar una simple repetición con `replicate()`.

```
# randomize sample to create two arbitrary samples
# replicate a random draw from a uniform distribution
rnd1.s <- replicate(stotal, round(runif(1),0))
```

O por `sample()` *con retroceso*:

```
# or that way (sample with replacement)
rnd2.s <- sample(c(0,1), stotal, replace=TRUE)
```

O por `sample()` *sin retroceso* como permutación:

```
# or that way (permutation, sample without replacement)
rnd3.perm <- sample(stotal, stotal, replace=FALSE)
# see sorted
sort(rnd3.perm)
```

Se puede inscribir las tres soluciones en una tabla

```
group <- rep(c(0,1),each=npersample)
# rnd3.s <- data.frame(order(rnd3.perm, group),group)
rnd3.s <- group[order(rnd3.perm,group)]
# all together
rnd123.tab <- data.frame(rnd1.s, rnd2.s, rnd3.s)
head(rnd123.tab)
tail(rnd123.tab)
```

y comparar las probabilidades resultantes y las probabilidades esperadas:

```
> # check probs
> # zeros
> apply(rnd123.tab,2,function(x) abs(sum(x-1)/(stotal)))
rnd1.s    rnd2.s    rnd3.s
0.4833333 0.4666667 0.5000000
> # ones
> apply(rnd123.tab,2,function(x) sum(x)/(stotal))
rnd1.s    rnd2.s    rnd3.s
0.5166667 0.5333333 0.5000000
```

Lo último que falta es la posibilidad de especificar probabilidades sobre cómo deben distribuirse los valores aleatorios. Las realizaciones pueden comprobarse inmediatamente con las probabilidades generadas. Con muestras más grandes, el valor empírico se acerca cada vez más al valor dado.

```
# sample with different probs R-Code
probs <- c(0.25, 0.5, 0.25)
items <- c("1|low","2|middle","3|high")
data.frame(probs,items)
reps <- 1e5
rnd4.s <- sample(items, reps, replace=TRUE, prob=probs)
head(rnd4.s)
tail(rnd4.s)
prop.table(table(rnd4.s))
1-(prop.table(table(rnd4.s))/probs)
```

De este modo, se puede derivar una aleatorización razonable para la mayoría de los estudios con un esfuerzo justificable. El paquete R `randomizr`, por otra parte, permite una interesante agrupación de los datos iniciales a partir de la cual se puede extraer. En primer lugar, mostramos la aleatorización completa sobre la base de predefinidos porcentajes.

```
> # full randomization according to probs / percentages
> rand1 <- complete_ra(N=100, m=33)
> table(rand1)
rand1
0  1
67 33
> rand2 <- complete_ra(N=100, m=50)
> table(rand2)
rand2
0  1
50 50
```

Veamos ahora las agrupaciones. Simplificando, tomamos como agrupaciones las diez últimas letras minúsculas del alfabeto (empezando por la z y terminando por la q) y las replicamos en orden ascendente. Es decir, la primera letra aparece una vez, la segunda dos veces, la tercera tres veces, y así sucesivamente. La aleatorización se realiza según determinados criterios de distribución (= probabilidades). distribuidos en estas agrupaciones (= clusters).

```
# clusters
> clust.num <- rep((letters)[26:17], times=1:10)
> clust.num
[1] "z" "y" "y" "x" "x" "x" "w" "w" "w" "w" "v" "v" "v" "v" "v" "u"
[17] "u" "u" "u" "u" "u" "t" "t" "t" "t" "t" "t" "t" "s" "s" "s" "s"
[33] "s" "s" "s" "s" "r" "r" "r" "r" "r" "r" "r" "r" "r" "q" "q" "q"
[49] "q" "q" "q" "q" "q" "q" "q"
> table(clust.num)
clust.num
q r s t u v w x y z
10 9 8 7 6 5 4 3 2 1
> length(clust.num)
[1] 55
```

La función de R `cluster_ra` distribuye aleatoriamente las letras anteriores de forma que cada una de las condiciones se dé en una determinada proporción con respecto a las demás. En el ejemplo hay cuatro condiciones: un grupo de control, un grupo placebo y dos grupos de tratamiento. Estos cuatro grupos se dan en las proporciones ficticias de 2, 5, 2 y 1, lo que suma $2+5+2+1 = 10$ que debe corresponder al número de letras anteriores (= agrupaciones o clusters). Esto se presenta en forma de tabla y se calculan las sumas marginales para ver si las sumas son correctas.

```
rand2 <- cluster_ra(cluster=clust.num, m_each=c(2,5,2,1),
           conditions=c("control","placebo","treat1","treat2"))
```

Ahora comprobamos las distribuciones resultantes:

```
> table(rand2, clust.num)
clust.num
rand2   q  r s t u v w x y z
control  0 0 8 7 0 0 0 0 0 0
placebo 10 0 0 0 6 5 4 0 0 1
treat1   0 9 0 0 0 0 0 3 0 0
treat2   0 0 0 0 0 0 0 0 2 0
> # add margins
> addmargins(table(rand2, clust.num),c(1,2))
clust.num
rand2   q  r s t u v w x y z  Sum
control  0 0 8 7 0 0 0 0 0  0 15
placebo 10 0 0 0 6 5 4 0 0  1 26
treat1   0 9 0 0 0 0 0 3 0  0 12
treat2   0 0 0 0 0 0 0 0 2  0  2
Sum 10   9 8 7 6 5 4 3 2 1 55
> sum(1:10)
[1] 55
> sum(c(2,5,2,1))
[1] 10
> c(2,5,2,1)/sum(c(2,5,2,1))
[1] 0.2 0.5 0.2 0.1
> rowSums(table(rand2, clust.num))
control placebo treat1 treat2
15 26 12 2
> rowSums(table(rand2, clust.num))/sum(rowSums(table(rand2, clust.num)))
control    placebo    treat1     treat2
0.27272727 0.47272727 0.21818182 0.03636364
```

Dado el pequeño tamaño de la muestra, esto funciona bastante bien.Cambiemos el paquete R. El paquete R `randomizeR` está especialmente diseñado para ensayos clínicos con el fin de minimizar sesgos como el sesgo de selección (selection bias; Tripepi, Jager, Dekker & Zoccali, 2010) o para implementar diseños de bloques aleatorizados etc. con poco esfuerzo. El sesgo de selección, por ejemplo, crea distorsiones como resultado de una selección desfavorable de la muestra que no representa a la población objetivo, lo que supone una desviación sistemática de la aleatorización. Se pueden generar secuencias de cadenas aleatorias con `randomizeR`. Para muestras pequeñas con $N \leq 24$, estas secuencias se pueden emitir completamente, es decir, se generan todas las combinaciones. Para muestras más grandes, se generan secuencias de referencia con algunos miles de dígitos.

El primer paso es definir dos muestras con $n_1 = 100$ y $n_2 = 10$ elementos. A continuación, siga las llamadas para generar los parámetros para una aleatorización completa o un diseño de bloques permutados.

```
> n1 <- 100
> n2 <- 10
> # complete randomization
> params.cr <- crPar(n1)
> # permuted block randomization
> params.pbr <- pbrPar(n2)
> params.cr
Object of class "crPar"
design = CR
N = 100
groups = A B
> params.pbr
Object of class "pbrPar"
design = PBR(10)
bc = 10
N = 10
groups = A B
params.pbr
```

El valor de referencia de la secuencia a generar va seguido de las llamadas para generar las secuencias aleatorias según los parámetros ya seleccionados anteriormente.

```
# set reference size R-Code
rsize <- 1e4
# generate sequences
cr.seq <- genSeq(params.cr, r=rsize, seed=seed)
cr.seq
str(cr.seq)
pbr.seq <- genSeq(params.pbr, r=rsize, seed=seed)
pbr.seq
str(pbr.seq)
```

Creamos la secuencia completa con la muestra más pequeña n_1

```
# for small sizes complete is possible
# = power set count
pbr.all.seq <- getAllSeq(params.pbr)
pbr.all.seq
str(pbr.all.seq)
pbr.all.seq@M
cr.all.seq <- getAllSeq(crPar(n2))
cr.all.seq
str(cr.all.seq)
cr.all.seq@M
getRandList(pbr.all.seq)
```

– a modo de impresión las dimensiones

```
> dim(pbr.all.seq@M)
[1] 252 10
> dim(cr.all.seq@M)
[1] 1024 10
```

e insertamos una función R separada que produzca la secuencia con su probabilidad.

```
seqAprobs <- function(seque)
{
seqscollapse <- apply(getRandList(seque),1,
function(x) paste(x,collapse=""))
ps <- getProb(seque)
return(data.frame(sequences=seqscollapse, probs=ps))
}
# call:
# seqAprobs(seque=pbr.all.seq)
```

con la llamada

```
seqAprobs.pbr <- seqAprobs(seque=pbr.all.seq)
```

Como estas secuencias completas son bastante largas, basta con ver el principio y el final para hacerse una idea.

```
> head(seqAprobs.pbr)
    sequences  probs
1 BBBBBAAAAA 0.003968254
2 BBBBABAAAA 0.003968254
3 BBBABBAAAA 0.003968254
4 BBABBBAAAA 0.003968254
5 BABBBBAAAA 0.003968254
6 ABBBBBAAAA 0.003968254
> tail(seqAprobs.pbr)
```

```
     sequences   probs
247 BAAAAABBBB 0.003968254
248 ABAAAABBBB 0.003968254
249 AABAAABBBB 0.003968254
250 AAABAABBBB 0.003968254
251 AAAABABBBB 0.003968254
252 AAAAABBBBB 0.003968254
```

Una característica especial de `randomizeR` es la posibilidad de comparar diferentes secuencias aleatorias si se da un cierto criterio. Las `corGuesses`, es decir, las suposiciones correctas, son un ejemplo de dicho criterio o también del sesgo de selección, véase `?compare`. Las `corGuesses` representan los intentos correctos de adivinar la probabilidad de la secuencia de aleatorización basada en la probabilidad de conjetura. Las secuencias completas anteriores de las dos muestras pequeñas sirve como base de datos para la aleatorización completa `cr.all.seq`. y para el diseño de bloques permutados `pbr.all.seq`.

```
> comp.res0 <- compare(issue=corGuess("CS"), cr.all.seq, pbr.all.seq)
> comp.res0
Comparison for propCG(CS)
CR PBR.10.
mean 0.500 0.653
sd 0.137 0.060
max 0.750 0.750
min 0.050 0.550
x05 0.250 0.550
x25 0.400 0.600
x50 0.500 0.650
x75 0.600 0.700
x95 0.700 0.750
```

Para la comparación con el sesgo de selección, es necesario especificar los valores esperados 0 o 1 de las variables dependientes de los dos grupos de tratamiento (o grupo de tratamiento y grupo de control), o similares) y las correspondientes desviaciones estándar σ_0 ó σ_1. Además, existen supuestos sobre el tamaño del efecto y la tasa de error de tipo I. A efectos de demostración, basta con valores ficticios de $\mu_0 = 2$, $\mu_1 = 3$, $\sigma_0 = 1$, $\sigma_1 = 1$, $\eta = 0.2$ y $\alpha = 0.03$. Esto supone que los dos grupos de tratamiento difieren. Eso no tiene ser así en la práctica.

```
> # expected responses and standard deviations in
> # treatment groups of clinical trials
> endpoints <- normEndp(mu=c(2,3), sigma=c(1,1))
> comp.res1 <- compare(issue=selBias(type="CS", method="exact",
           + eta=0.2, alpha=0.03), cr.all.seq,
           + pbr.all.seq, endp=endpoints)
> comp.res1
Comparison for P(rej)(CS)
     CR    PBR.10.
mean 0.164 0.162
sd   0.021 0.016
max  0.214 0.192
min  0.000 0.138
x05  0.133 0.138
x25  0.154 0.148
x50  0.167 0.161
x75  0.175 0.174
x95  0.192 0.192
```

Esto puede representarse gráficamente como un gráfico de violines o de cajas (véase la Fig. 4.48 a la izquierda para `comp.res0`, a la derecha para `comp.res1`).

```
plot(comp.res0)
plot(comp.res1)
plot(comp.res0, y="boxplot")
plot(comp.res1, y="boxplot")
```

Encontrará más detalles sobre las comparaciones de los procedimientos de aleatorización en Uschner (2018-15-06) o las páginas del manual y las viñetas sobre `randomizeR`, por ejemplo sobre el *desirability score* (valor de deseabilidad; Derringer & Suich, 1980), un índice que permite comparar procedimientos de aleatorización en base a criterios de optimización.

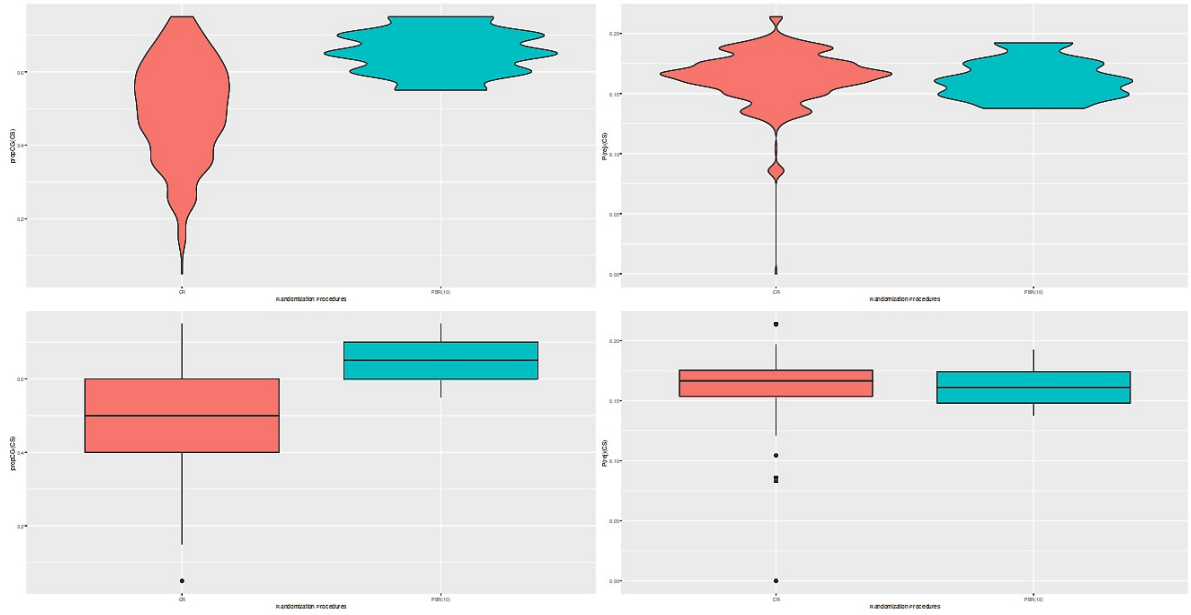


Figura 4.48. *Aleorización*

4.4.8 Datos faltantes

Los valores faltantes son un problema básico en estadística (Little & Rubin, 1987; Schafer, 1997), especialmente cuando no está claro por qué faltan los datos, cuántos faltan, pero se supone que las conclusiones, así como las predicciones, se basan exactamente en estos datos perdidos. La estadística clásica distingue tres casos de datos ausentes, cada uno de los cuales debe tratarse de forma diferente. Básicamente, se trata de entender si los datos faltan "por casualidad" o si hay un sistema detrás. ¿Se trata de un planteamiento sistemático basado en el objeto de investigación, de modo que pueda observarse una confusión entre el mecanismo que genera los datos y el mecanismo que conduce a la falta de datos?

Así, en un estudio a largo plazo sobre el desarrollo intelectual y emocional de los niños, podría ocurrir que precisamente quedaran fuera de la muestra los niños con menor inteligencia cuyos padres no tuvieran ni la capacidad y/o ni la voluntad ni la perspicacia para ocuparse del desarrollo de sus hijos en mayor medida de lo que ya lo hacen. El contacto con el estudio y las repetidas pruebas hacen que estos problemas (por ejemplo, la mala conciencia, las emociones negativas, la sensación de no poder hacer frente a la situación, etc.) salgan aún más a la superficie. El resultado es que estas familias y sus hijos quedan cada vez más fuera de la muestra. Se produce así un conflicto entre el el propio estudio y su tema de la inteligencia y las emociones, y que una proporción significativa de los niños que abandonan la muestra lo hacen por motivos relacionados con la inteligencia y las emociones en un sentido más amplio. Podría ser que algunos padres invirtieran más en sus hijos sin el estudio y que se obtuvieran resultados completamente diferentes. Pero no lo sabemos. Sería diferente si las familias se mudaran y, por tanto, vivieran fuera de la zona del estudio y también dejaran de formar parte de la muestra. En la medida en que los motivos del alejamiento no tengan nada que ver con la inteligencia de los niños o el clima emocional de las familias, se trataría de un caso de "azar", sea cual sea su definición (véase el capítulo 4.2.1 sobre el concepto de probabilidad).

Las causas se juzgan según el grado de aleatoriedad frente al de sistematicidad, siendo la aleatoriedad la conexión sistemática de las causas por la ausencia de datos y el objeto de investigación; se operativiza:

- *Falta completamente al azar* (missing completely at random; MCAR): no existe relación entre las causas de baja de una muestra y el objeto de estudio y sus variables implicadas. Este es el caso no problemático y todas las estimaciones son insesgadas. No se necesitan correcciones, ya que en el supuesto la suma de las distintas entradas da cero.
- *Falta al azar* (missing at random; MAR): existe una relación, pero las razones pueden dilucidarse completamente. Esta aclaración es de naturaleza sustantiva y no estadística. Las estadísticas no pueden explicar nada en este caso. Por ejemplo, la disposición a participar en un estudio clínico con cuestionario sobre problemas de salud mental podría ser menor en los hombres que en las mujeres. Pero estas razones no están directa o indirectamente relacionadas con el tema del cuestionario clínico, por ejemplo, depresión, ansiedad, etc. No obstante, pueden producirse problemas estadísticos si las celdas de diseño están infrapobladas en comparación con otras celdas. No obstante, las estimaciones siguen siendo insesgadas.
- *Falta no aleatoria* (missing not at random; MNAR): existe una relación entre el mecanismo que genera los datos y el responsable del abandono de la muestra. Más arriba se ha dado un ejemplo. Esto es problemático y las estimaciones están sesgadas. Una posibilidad sería introducir más variables en el modelo que puedan representar este mecanismo.

Tabla 4.48. *Tratmiento de datos faltantes*

NA Proceso	Anulación por listas		Imputación	
	Sesgo	*Varianza*	*Sesgo*	*Varianza*
MCAR	ninguno	alta	ninguno	reducida
MAR	Sí	alta	ninguno	reducida
MNAR	Sí	alta	reducido	reducida

Se pueden identificar distintas variantes para tratar los datos que faltan.

- La *supresión de conjuntos de datos* (por listas o por pares) pertenece sin duda al ámbito de "tirar información" y debe evitarse en la medida de lo posible y de todas las formas imaginables. Reducir aún más una muestra que ya es demasiado pequeña es una mala solución para un problema que puede abordarse de otra manera. puede enfocarse de otra manera.
- Sustituir los valores que faltan *por ceros* es legítimo, en el mejor de los casos, si los datos pueden tener realmente un contenido cero y los valores que faltan se codifican como ceros. Pero entonces no se trata realmente de valores perdidos si, por ejemplo, las personas responden a la pregunta de cuánto alcohol beben a la semana con "nada" (= cero) y esto se codifica como NA. Cualquier otra cosa causa distorsiones y burdas tonterías si, por ejemplo, una variable no puede ser cero por razones de plausibilidad – por ejemplo, valores de CI, mediciones fisiológicas, etc.
- *Sustituir* los valores que faltan por *la moda o la media* es una solución bastante difícil y debe evitarse. La razón es que aquí se favorece un determinado valor único y, por tanto, todas las desventajas, como la sensibilidad a los valores atípicos, salen a la luz y surten efecto en el caso del valor medio. El resultado son correlaciones y coeficientes de regresión distorsionados en las estimaciones del modelo.
- Lo mismo se aplica a la idea de trabajar con *regresiones como sustitutos*, que ya no incluyen todos los datos originales y, por lo tanto, también desperdician información valiosa o provocan distorsiones, a menos que (véase más adelante) se utilice otro algoritmo de optimización en el contexto de todos los datos disponibles.

- Si esto no se tiene en cuenta, la sustitución por regresión conduce a una sobreestimación de las covarianzas, es decir, de las correlaciones con otras variables.
- La *imputación en caliente* (hot-deck imputation) utiliza nuevos valores a partir de casos existentes, pero tiende a producir coeficientes y errores estándar relativamente sesgados. Además, aquí se duplican los valores sin razones más profundas y, por tanto, se favorecen rangos de valores sin que esto esté ampliamente justificado desde el punto de vista estadístico o de contenido. El resultado es una distorsión de los análisis siguientes.
- La *imputación simple o múltiple* (Rubin, 1987, 1996; Yang, 2016-02) simula los datos a nivel de persona y crea así una incertidumbre adicional, lo que aumenta las varianzas y, por tanto, la precisión. En el caso de la imputación múltiple, las imputaciones individuales deben agregarse cada una por agregada ("pooled") para cada fecha que falte a lo largo de los análisis realizados. La imprecisión resultante es naturalmente mayor que con las imputaciones simples. Se dispone de múltiples estimaciones simuladas que, a la inversa, aumentan la robustez de la estimación. Para cada imputación todos los análisis de datos se realizan por separado; y puesto que los análisis generados de este modo se agrupan posteriormente, en la práctica esto se puede hacer para cada ejecución de imputación. En la práctica, esto puede dar lugar a cálculos complicados, dependiendo del procedimiento de análisis. Van Buuren y Groothuis-Oudshoorn (2011, Fig. 1, p.5) visualizan el proceso y dan (ibid., S.6) una visión de los problemas en el contexto de las imputaciones múltiples.
- Se puede intentar modelizar el mecanismo que conduce a la falta de datos. Esto puede incorporarse al modelo general, por ejemplo, mediante análisis de clases latentes. Este enfoque puede ser una forma legítima de abordar el problema. Sin embargo, esto requiere un análisis preciso de cómo funciona el mecanismo en detalle, de lo contrario se producirán distorsiones.
- En el *emparejamiento predictivo de significados* (Gaert, Meinfelder & Bosch, 2016-01-25) se estima un modelo de regresión lineal en el que las variables con valores perdidos se modelan como variables dependientes. Para la estimación del modelo solo se toman los valores existentes. A partir del modelo estimado de este modo se extraen valores aleatorios para crear un nuevo conjunto de ponderaciones de regresión. A partir de ahí, se calculan las predicciones para los valores observados y ausentes con el fin de extraer aleatoriamente los valores que mejor se ajusten para aquellos casos en los que se observen los valores.
- En consecuencia, las desviaciones sumarias son mínimas, pero los valores individuales no se ajustan necesariamente. El procedimiento suele combinarse con la imputación múltiple y conduce a las mejores y más sólidas estimaciones para los valores que faltan a nivel de grupo.
- Los algoritmos de optimización para estimar la máxima verosimilitud (maximum likelihood) utilizan, por ejemplo, el algoritmo de maximización de expectativas (*Expectation Maximization;* EM) (Dempster, Laird & Rubin, 1977) o el de máxima verosimilitud con información completa (*Full Information Maximum Likelihood*; FIML). Con FIML, sólo se estiman los parámetros (por ejemplo, las covarianzas/varianzas), no los datos que faltan a nivel de persona, como ocurre con la imputación. Se toman los valores que parecen más plausibles para la muestra en cuestión. Como subproducto, se obtienen errores estándar para poder estimar el aumento de la incertidumbre ante los datos que faltan. El algoritmo EM, a su vez, funciona en dos pasos: En primer lugar, los valores que faltan se estiman a partir de los valores disponibles mediante regresión. Para minimizar el sesgo que surge en la verosimilitud se maximiza iterativamente en el segundo paso hasta que la solución es estable y el algoritmo de estimación converge. La solución resultante puede entenderse como el óptimo en las condiciones dadas.

Los datos ausentes son un problema primordial en la estadística clásica, ya que afectan a la distribución muestral de las estadísticas de interés o la visión básica del problema (Jaynes, 2003; p.534). Los procedimientos de estimación anteriores son muy propensos a errores, ya que la estimación en sí nunca es única (Little, 1988; Little & Rubin, 1983; Little & Rubin, 1987). En cambio, en la estadística bayesiana *no existe un equivalente* a los datos que faltan. En se toma la información (datos) de que se dispone. Por supuesto, que falten datos significa que falta información. Pero es así. Los algoritmos no cambian por ello, sólo porque los datos difieran, ya que según Jaynes (ibid., p.533)

> „One can write a single computer program which, once and for all, accepts whatever data (that is, whatever set of numbers $\{x_i; t_i\}$) we give it, and proceeds to do the correct calculations for that data set."

Esto significa que en la estadística bayesiana no se distingue entre datos que faltan y parámetros desconocidos. El proceso de inferencia al que se refiere Jaynes para llegar de los datos a las estimaciones se extiende de forma casi automática a los modelos con datos que faltan. Esto requiere un modelo apropiado que describa los datos observados o ausentes y los parámetros del modelo juntos en una distribución conjunta („joint distribution"). La estimación del modelo se lleva a cabo como de costumbre mediante MCMC (véase el capítulo 6.13). Una ventaja es que la información contextual conocida sobre la aparición de los datos que faltan puede utilizarse como información previa (prior information) y no es necesario excluir los casos con datos parcialmente disponibles. Estos supuestos del modelo están justificados teóricamente en el marco de la estadística de Bayes, son coherentes internamente con respecto a la estimación del modelo calculado y permiten expresar la incertidumbre encontrada en el curso de los datos que faltan. No obstante, la cuestión de los datos que faltan también se debate e investiga en el ámbito bayesiano, precisamente porque los mecanismos mencionados que subyacen a los datos que faltan suelen ser muy poco claros en la práctica. Si un mecanismo de este tipo crea datos sesgados, es irrelevante qué variante estadística o qué método se utilice para la evaluación, ya que el sesgo no puede eliminarse. Esto hace aconsejable distinguir entre variables dependientes e independientes que faltan y mecanismos ignorables y no ignorables. Así pues, hay que distinguir entre "¿qué falta?" y "¿cómo se produce realmente la ausencia?" Siempre resulta difícil cuando faltan muchas covariables. Cuantos menos valores faltantes haya en las covariables, mejor.

Dada la importancia del tema, R tiene su propia página de resumen (CRAN, 2019c) que enumera los paquetes R relevantes. Mächler (2015) mantiene otra lista extensa. Entre otros, cabe destacar los paquetes de `R sjmisc, mi, VIM, mice, Hmisc, Amelia, mitools, pan` y `JointAI`. La atención se centra en la descripción y exploración de datos faltantes, métodos basados en la verosimilitud, imputación simple y múltiple, así como tipos de datos especiales y campos de aplicación. Los paquetes de R más especializados, como `RSiena` para el análisis de redes sociales tienen sus propios mecanismos para sustituir los datos que faltan (Krause, 2019) o recurren como el paquete R `brms` mediante `brm_mutiple()` a otros paquetes R como `mice` (Bürkner, 2019-05-23). Tutoriales (por ejemplo, Analytics Vidhya Content Team, 2016; Leeper, 2009), viñetas de R (por ejemplo, Erler, 2019-06-06; Gelman, Hill, Su, Yajima, Pittau, Goodrich, Si y Kropko, 2015-06-16; Ripley, Snijders, Boda ,Vörös y Preciado, 2019-05-21) y una amplia variedad de tutoriales en blogs (por ejemplo, Allison, 2015; Rickert, 2016; Joachim, s.f.) sobre diferentes procedimientos analíticos completan el panorama, por no mencionar capítulos de libros relevantes (por ejemplo, Gelman & Hill, 2007, cap. 25) y artículos de revistas (Honaker, King & Blackwell, 2011; Su, Gelman, Hill & Yajima, 2011; Grund, Simon, Lüdtke, Oliver & Robitzsch, Alexander, 2016; Krause, Huisman & Snijders, 2018).

Azur, Stuart, Frangakis y Leaf (2011) ofrecen un análisis algo más antiguo de la imputación múltiple. Los autores señalan problemas en el contexto de la imputación múltiple, como una menor fundamentación teórica, posibles desviaciones de los supuestos de normalidad multivariante o posible agrupación de los datos, todos ellos puntos que pueden tener un impacto negativo. Esta no es la única razón por la que los paquetes de R como `mice, miceadds, micecmd, missMDA, smcfcs, mi`, etc. disponen ahora de una amplia gama de algoritmos de imputación que intentan proporcionar una imputación de datos faltantes para diferentes tipos de datos y escenarios de investigación (van Buuren, 2018). Los algoritmos utilizados se basan tanto en la estadística clásica como en la bayesiana para estimar los datos que faltan.

La tabla 4.10 resume de forma muy simplificada el problema de la eliminación de conjuntos de datos con datos faltantes frente a la imputación de datos faltantes. A modo de orientación, puede observarse que con un diagnóstico cuidadoso de las causas de los datos faltantes y una aplicación cuidadosa, el enfoque de la imputación conduce prácticamente de forma sistemática a una mejora de la situación de los datos en comparación con la eliminación de conjuntos de datos debido a la falta de datos, es decir, no hacer nada. Por un lado, la imputación múltiple crea un cierto aumento de la varianza por la introducción de valores estimados, pero por otro, esta incertidumbre se reduce con la estimación o estabiliza las estimaciones, de modo que en conjunto las ventajas superan a los inconvenientes. La supresión de datos no da lugar a sesgos o los da de forma significativa, dependiendo del mecanismo subyacente de los datos que faltan. En cualquier caso, la varianza se altera de forma ambigua porque simplemente hay menos datos. Esto puede tener un efecto desfavorable, especialmente con muestras pequeñas. De poco sirve saber que con muestras muy

grandes y en el caso de datos que faltan, las estimaciones son correctas. Si no se conocen los valores de la población y por eso se suelen realizar estudios – y si las muestras son pequeñas, es realista esperar que se produzcan sesgos, pero no sabemos exactamente en qué dirección y en qué medida. Si lo supiéramos, no necesitaríamos investigar más.

Lo que pueden causar los datos que faltan si "sólo" los eliminamos se muestra con un sencillo ejemplo libre de contexto con la función R `na.sim()`. Tenemos una muestra aleatoria de tamaño *n* (distribución normal estándar) y eliminamos simplemente el 10 % de los datos como "datos que faltan". Para la muestra original y la reducida se calculan los estadísticos descriptivos y se comparan entre sí como la proporción de *datos originales/conjunto de datos reducidos* (`ptII_quan_classicstats_missingdata.r`).

```
na.sim <- function(n=1000, MW=0, SD=1, pct.rem=0.1, seed=09877) R-Code
{
set.seed(seed)
v <- rnorm(n=n, mean=MW, sd=SD)
v.descstat <- c(N=length(v),summary(v), var=var(v), sd=sd(v))
rem <- sample(1:n, size=n*pct.rem, replace=FALSE)
v.red <- v[-rem]
v.red.descstat <- c(N=length(v.red),summary(v.red),
           var=var(v.red), sd=sd(v.red))
res <- t(data.frame(v=v.descstat, v.red=v.red.descstat,
           v.ratio=v.descstat/ v.red.descstat))
return(res)
}
```

Lo llamamos con `na.sim()` sin más parámetros.

```
            N       Min.       1st Qu.     Median             Mean
v 1000.000000 -3.035912 -0.6486754 -0.008827667 -0.02660465
v.red 900.000000 -3.035912 -0.6488146 -0.011653540 -0.03526179
v.ratio 1.111111 1.000000 0.9997856 0.757509461 0.75448964
        3rd Qu.    Max.    var        sd
v       0.6091743 3.648211 0.9348516 0.9668773
v.red   0.5898252 3.648211 0.9271335 0.9628777
v.ratio 1.0328048 1.000000 1.0083248 1.0041538
```

Obviamente, las distintas estadísticas descriptivas difieren. La mediana y la media parecen mostrar una desviación similar a la original. La varianza y la desviación estándar han aumentado ligeramente y el mínimo y el máximo permanecen constantes. Sin embargo, esto no tiene por qué ser así, como ilustra la elección de un valor inicial diferente para el generador aleatorio.

```
seed <- 0987
na.sim(seed=seed)
```

Ahora las desviaciones entre la mediana y la media son drásticas, mientras que los cambios para la varianza y la desviación estándar son muy similares a los del primer ejemplo. Esto demuestra que, dependiendo de los valores omitidos y ausentes, las estadísticas se calculan sobre una base de datos diferente. Si faltan valores atípicos, la media cambia mucho, la mediana menos y para las varianzas depende. Sin embargo, al aumentar el tamaño de la muestra y el MCAR real esta desviación se hace menor.

```
na.sim(n=1e6, seed=seed)
na.sim(n=1e7, seed=seed)
```

Aquí es sólo un vector y, por tanto, sólo faltan puntos de datos. Si se tratara de una matriz, es decir, muchas variables, se perderían filas enteras y el tamaño de la pérdida de datos se multiplica considerablemente por la eliminación de los conjuntos de datos con datos que faltan. Dependiendo de las estadísticas descriptivas, aquí se producen cambios más o menos grandes, que, además, están sujetos a los efectos aleatorios pertinentes de la generación de la muestra. Es fácil imaginar lo que esto significa para las pruebas estadísticas inferenciales derivadas de conjuntos de datos más grandes.

En otro ejemplo, se utilizan los paquetes de R `mice, Amelia, brms` y `mi`. Comparamos los resultados respectivos entre sí en un ejemplo empírico, el conocido conjunto de datos Iris (Anderson, 1935, 1936, véase el Cap. 5.5.3 para un uso adicional de los datos Iris). El conjunto de datos contiene las variables longitud y anchura de sépalo y pétalo para n = 50 flores de tres variantes de la especie Iris (Iris setosa, Iris versicolor, Iris virginica). El conjunto de datos está completo en el original y no contiene datos que falten. Éstos se generan artificialmente para que los métodos de imputación puedan probarse en el original como referencia. Como problema estadístico, se utiliza el siguiente modelo para predecir la longitud de los sépalos en todas las especies a partir de todas las variables. En la notación habitual de R es (`ptII_quan_-classicstats_missingdata.r`).

```
Sepal.Length ~ Sepal.Width + Petal.Width + Petal.Length + Species
```

Se utiliza un modelo lineal sencillo `lm()`:

```
# use iris data
data(iris)
head(iris)
tail(iris)
dim(iris)
# check whether there are NAs
which(is.na(iris))
# linear model
iris.lm <- lm(Sepal.Length ~ Sepal.Width + Petal.Width +
           Petal.Length + Species, data=iris)
iris.lm.res <- summary(iris.lm)$coefficients
```

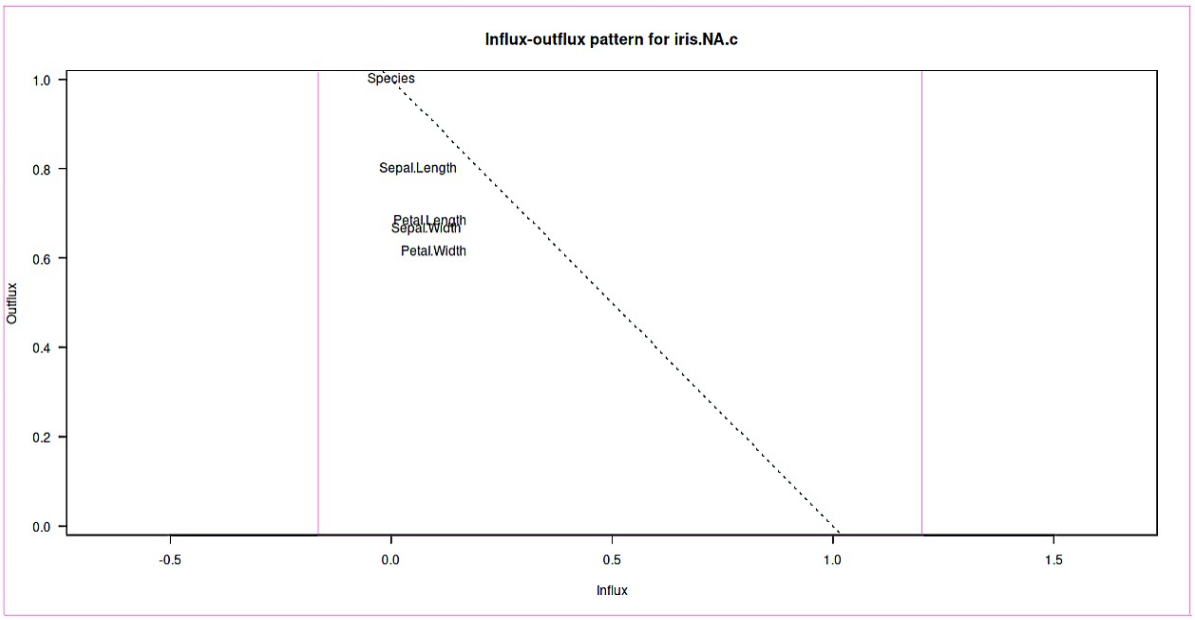


Figura 4.49. *Datos que faltan en el conjunto de datos del iris (patrón NA con `fluxplot()`)*

El modelo lineal resulta en

```
Call: R-Output
lm(formula = Sepal.Length ~ Sepal.Width + Petal.Width +
Petal.Length + Species, data = iris)
Residuals:
 Min      1Q      Median  3Q      Max
-0.79424 -0.21874 0.00899 0.20255 0.73103
Coefficients:
                 Estimate  Std. Error  t value Pr(>|t|)
```

```
(Intercept)        2.17127 0.27979    7.760  1.43e-12 ***
Sepal.Width        0.49589 0.08607    5.761  4.87e-08 ***
Petal.Width       -0.31516 0.15120   -2.084  0.03889 *
Petal.Length       0.82924 0.06853   12.101  < 2e-16 ***
Speciesversicolor -0.72356 0.24017   -3.013  0.00306 **
Speciesvirginica  -1.02350 0.33373   -3.067  0.00258 **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.3068 on 144 degrees of freedom
Multiple R-squared: 0.8673, Adjusted R-squared: 0.8627
F-statistic: 188.3 on 5 and 144 DF, p-value: < 2.2e-16
```

Ahora se trata de crear los datos que faltan – se elimina el 10% de los datos:

```
# percent to delete = 10% = create NAs
pc <- 0.1
# data points
datps <- prod(dim(iris[,1:4]))
seed <- 464645577
set.seed(seed)
del.IDs <- sample(datps, datps*pc, replace=FALSE)
del.IDs
iris.m <- as.matrix(iris[,1:4])
iris.m
iris.m[del.IDs] <- NA
iris.NA <- iris.m
iris.NA
```

y se exploran inmediatamente los NA y se muestran los datos gráficamente (véase la Fig. 4.49, para la función `fluxplot()` y la Fig. 4.49 para `md.pattern()`).

```
NA.IDs <- which(is.na(iris.NA), arr.ind=FALSE)
# check whether the proper values were deleted (=NA)
NA.IDs == sort(del.IDs)
iris.NA.c <- data.frame(iris.NA, Species=iris[,5])
fluxplot(iris.NA.c)
flux(iris.NA.c)
md.pairs(iris.NA.c)
md.pattern(iris.NA.c)
```

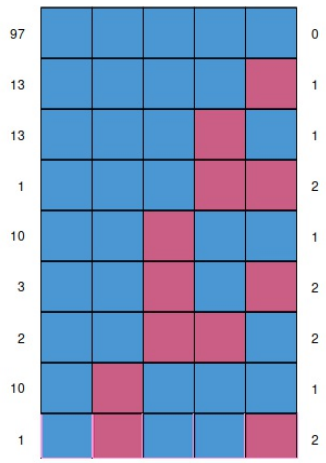


Figura 4.50. *Datos que faltan en el conjunto de datos del iris (patrón NA con `md.pattern()`)*

Pasemos ahora a las imputaciones y las estimaciones del modelo lineal en ausencia de datos. Comenzamos con el paquete de R `mice`, que trabaja con el ajuste predictivo de medias (predictive mean

matching) por defecto y genera cinco conjuntos de imputación. Luego se calcula el modelo lineal para cada conjunto y se agrupan los resultados para tener en cuenta las mayores varianzas generadas por los conjuntos de imputación (van Buuren y Groothuis-Oudshoorn, 2011).

```
# imputation via mice
# pmm = predictive mean matching
iris.imp <- mice(iris.NA.c, method=c("pmm"))
iris.imp.fit <- lm.mids(Sepal.Length ~ Sepal.Width + Petal.Width +
                  Petal.Length + Species, data=iris.imp)
iris.imp.fit
iris.imp.mice.pooled <- pool(iris.imp.fit)
summary(iris.imp.mice.pooled)
```

La salida con summary() contiene lo siguiente:

```
    term             estimate  std.error  statistic  df     p.value
1 (Intercept)        2.3477979 0.29788918  7.881447  62.718 5.914536e-11
2 Sepal.Width        0.4492698 0.09162514  4.903347  55.931 8.477026e-06
3 Petal.Width       -0.4420914 0.19418848 -2.276610  28.239 3.058303e-02
4 Petal.Length       0.8394525 0.08041792 10.438625  34.996 2.727818e-12
5 Speciesversicolor -0.6892232 0.24258991 -2.841104 110.548 5.353932e-03
6 Speciesvirginica  -0.8179694 0.34031365 -2.403575 104.193 1.800256e-02
```

El script de R contiene unas cuantas llamadas más para comparar diferentes soluciones estableciendo los coeficientes de las diferentes soluciones en relación a la solución con `mice()`. Aquí, `iris.lm.res` es el resultado sin valores perdidos e `iris.lm.NA.res` es el que tiene valores perdidos, en función de las estimaciones del modelo lineal.

```
# compare as ratio R-Code
# just estimates
iris.lm.res[,"Estimate"]/iris.imp.mice.pooled$pooled[,c("estimate")]
iris.lm.NA.res[,"Estimate"]/iris.imp.mice.pooled$pooled[,c("estimate")]
# compare all as ratio
iris.lm.res/summary(iris.imp.mice.pooled)[,c("estimate","std.error",
           "statistic","p.value")]
iris.lm.NA.res/summary(iris.imp.mice.pooled)[,c("estimate","std.error",
           "statistic","p.value")]
```

Ahora sigue `Amelia` (Honaker, King & Blackwell, 2018-05-07), que también incluye instrumentos para explorar los datos que faltan y, por lo demás, recuerda mucho a `mice` en cuanto al procedimiento, a saber, imputación (múltiple), cálculo y agrupación de los coeficientes y errores estándar más salida como gráficos (véanse las Fig. 4.51 y 4.52). Esto puede ir seguido, como en el caso anterior de `mice()`, de comparaciones con las estimaciones sin valores omitidos o con valores omitidos y sobre la base del modelo lineal simple.

```
# via Amelia
# https://www.r-bloggers.com/how-to-perform-a-logistic-regression-in-r/
missmap(iris.NA.c)
iris.ame <- amelia(iris.NA)
str(iris.ame)
summary(iris.ame)
plot(iris.ame)
le <- length(iris.ame$imputations)
betas <- vector()
ses <- vector()
for(i in 1:le)
{
dats <- data.frame(iris.ame$imputations[[i]], Species=iris$Species)
iris.ame.lm <- lm(Sepal.Length ~ Sepal.Width + Petal.Width +
           Petal.Length + Species, data=dats)
betas <- rbind(betas, iris.ame.lm$coef)
ses <- rbind(ses, coef(summary(iris.ame.lm))[,2])
```

```
}
betas
ses
iris.imp.amelia.pooled <- mi.meld(q=betas, se=ses)
iris.imp.amelia.pooled
iris.imp.amelia.pooled.tab <- t(do.call("rbind",iris.imp.amelia.pooled))
colnames(iris.imp.amelia.pooled.tab) <- c("Estimate","Std. Error")
iris.imp.amelia.pooled.tab
```

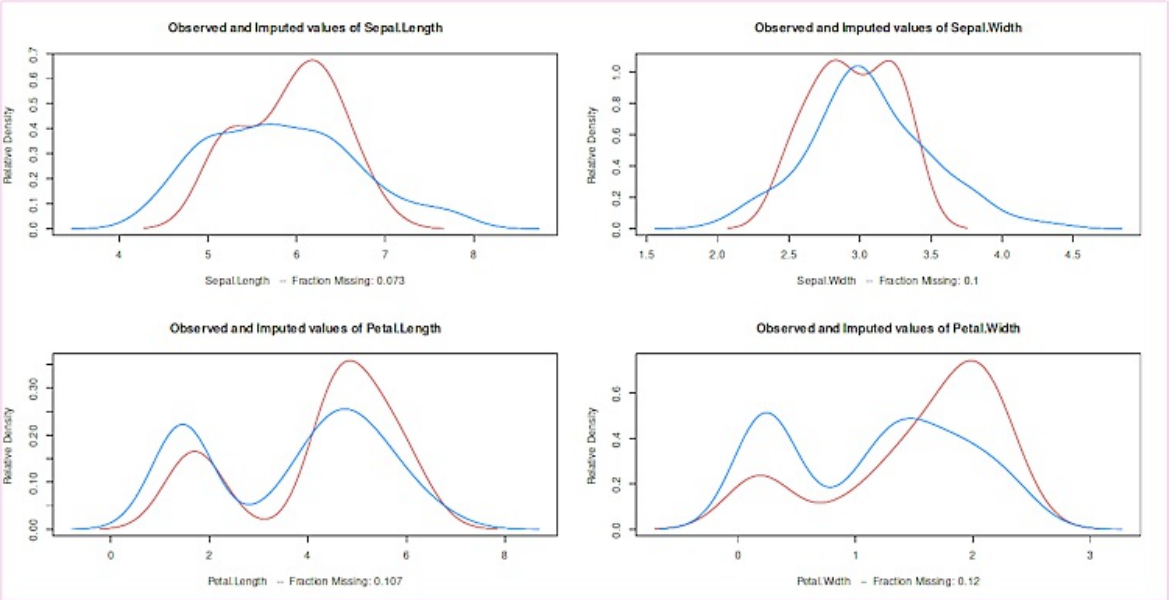


Figura 4.51. *Datos que faltan (conjunto de datos del iris, método* `Amelia`*, patrón NA).*

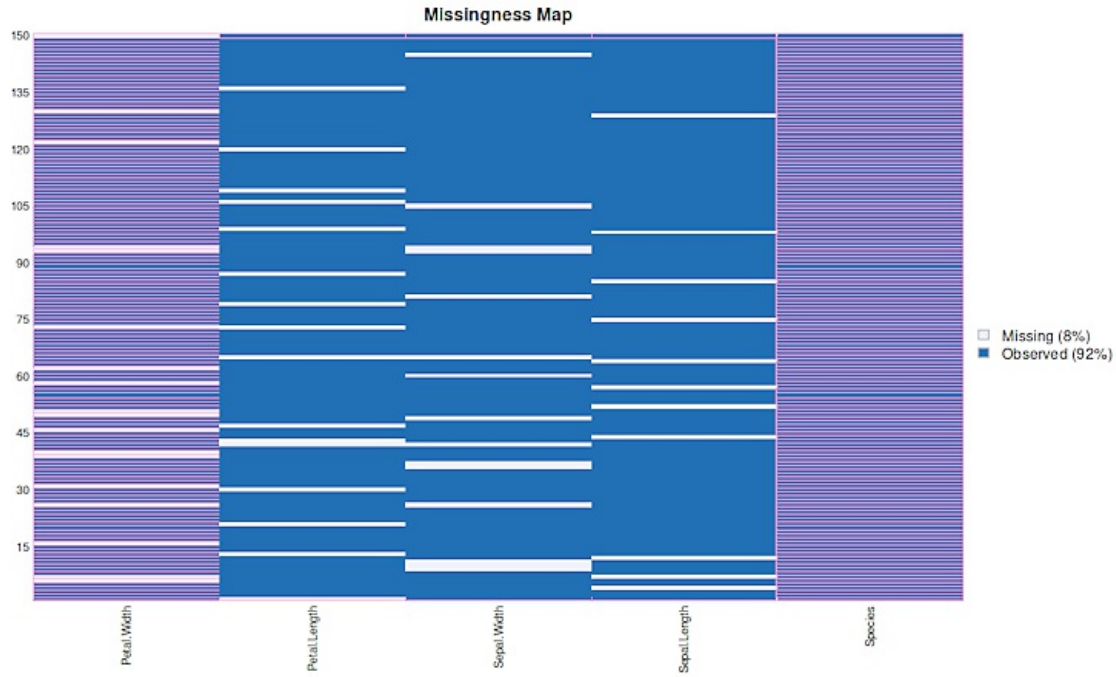


Figura 4.52. *Datos faltantes (conjunto de datos Iris, método* `Amelia`*, estimación de valores NA)*

Una variante bayesiana también utiliza `mice` (Bürkner, 2019-05-23), pero calcula un modelo Bayes lineal. La función de R `brm_multiple()` incluye tanto el paso de imputación como la estimación del modelo lineal.

```
# Bayes
# https://cran.r-project.org/web/packages/brms/vignettes/brms_missings.html
#
# gives out warnings with Rhats >1 (we just ignore that...)
# in reality one should consider better priors, more iterations, etc.
iris.imp.brmsfit <- brm_multiple(Sepal.Length ~ Sepal.Width +
           Petal.Width + Petal.Length +
           Species, data=iris.imp, chains=2)
# does not give a warning
iris.brmsfit <- brm(Sepal.Length ~ Sepal.Width + Petal.Width +
           Petal.Length + Species, data=iris, chains=2)
summary(iris.imp.brmsfit)
plot(iris.imp.brmsfit, pars = "^b_")
iris.imp.brmsfit$rhats
fe.iris.imp.brms <- fixef(iris.imp.brmsfit)
fe.iris.imp.brms
```

Los resultados gráficos son numerosos y los omitimos aquí. La estimación múltiple final es

```
Family: gaussian
Links: mu = identity; sigma = identity
Formula: Sepal.Length ~ Sepal.Width + Petal.Width +
    Petal.Length + Species
Data: iris.imp (Number of observations: 150)
Samples: 10 chains, each with iter = 2000; warmup = 1000; thin = 1;
total post-warmup samples = 10000
Population-Level Effects:
                  Estimate Est.Error l-95% CI u-95% CI
Intercept          2.35     0.30      1.75     2.93
Sepal.Width        0.45     0.09      0.27     0.63
Petal.Width       -0.44     0.19     -0.79    -0.05
Petal.Length       0.84     0.08      0.68     0.99
Speciesversicolor -0.69     0.24     -1.16    -0.22
Speciesvirginica  -0.82     0.34     -1.49    -0.16
                  Rhat  Bulk_ESS Tail_ESS
Intercept         1.06  107       809
Sepal.Width       1.07   89       422
Petal.Width       1.15   45       130
Petal.Length      1.12   52       492
Speciesversicolor 1.03  303      3553
Speciesvirginica  1.04  260      1372
Family Specific Parameters:
       Estimate Est.Error l-95% CI  u-95% CI  Rhat  Bulk_ESS Tail_ESS
sigma  0.31     0.02      0.27      0.35      1.15  45       197

Samples were drawn using sampling(NUTS). For each parameter,
Bulk_ESS and Tail_ESS are effective sample size measures,
and Rhat is the potential scale reduction factor on split
chains (at convergence, Rhat = 1).

Warning:
Parts of the model have not converged (some Rhats are > 1.05).
Be careful when analysing the results! We recommend running
more iterations and/or setting stronger priors.
```

El mensaje de advertencia primero dice que nuestro modelo todavía no es realmente bueno y que deberíamos elegir mejor los valores a priori. Esta advertencia no viene con una simple llamada, porque:

```
> summary(iris.brmsfit)
Family: gaussian
Links: mu = identity; sigma = identity
Formula: Sepal.Length ~ Sepal.Width + Petal.Width +
           Petal.Length + Species
           Data: iris (Number of observations: 150)
Samples: 2 chains, each with iter = 2000; warmup = 1000; thin = 1;
total post-warmup samples = 2000
```

```
Population-Level Effects:
                  Estimate Est.Error l-95% CI u-95% CI
Intercept          2.17     0.30       1.60      2.74
Sepal.Width        0.50     0.09       0.31      0.68
Petal.Width       -0.31     0.16      -0.61     -0.01
Petal.Length       0.83     0.07       0.69      0.96
Speciesversicolor -0.71     0.24      -1.19     -0.24
Speciesvirginica  -1.01     0.34      -1.67     -0.35
                  Rhat Bulk_ESS Tail_ESS
Intercept         1.00 1399     1323
Sepal.Width       1.00 1065     1227
Petal.Width       1.00 1275     1137
Petal.Length      1.00 1313     1290
Speciesversicolor 1.00  688     1042
Speciesvirginica  1.00  696     1044
Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma 0.31     0.02      0.28     0.35     1.00 1722     1297
Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

La página de ayuda de `brm_multiple()` lo explica con más detalle:

```
Details                                                                R-manpage
The combined model may issue false positive convergence warnings,
as the MCMC chains corresponding to different datasets may not
necessarily overlap, even if each of the original models did
converge. To find out whether each of the original models converged,
investigate fit$rhats, where fit denotes the output of brm_multiple.
```

Con el paquete R `future`, se pueden utilizar varios núcleos de CPU para cálculos más complejos.

```
# parallelization
plan(strategy="multicore")
```

A continuación, se realizan las llamadas anteriores con `brm_multiple()`. Por último, el paquete de R `mi`, que también trabaja con imputación múltiple y lo hace dentro de un marco bayesiano (Goodrich & Kropko, 2014-06-16). Dejamos de lado el necesario trabajo preparatorio sobre las distribuciones a priori pero recordemos que esto es esencial en el contexto de Bayes (véase la sección 6.12). `mi` contiene algunas herramientas de diagnóstico, incluida una sencilla pero elegante comparación gráfica del conjunto de datos original con el imputado. También se ofrecen interesantes de diagnóstico. Por lo demás, el procedimiento se corresponde con las explicaciones anteriores.

```
iris.NA.c
mdf <- missing_data.frame(iris.NA.c)
show(mdf)
image(mdf)
hist(mdf)
iris.imp.mi <- mi(mdf, n.iter=30, n.chains=4, max.minutes=20)
show(iris.imp.mi)
round(mipply(iris.imp.mi, mean, to.matrix = TRUE), 3)
Rhats(iris.imp.mi)
plot(iris.imp.mi)
image(iris.imp.mi)
summary(iris.imp.mi)
iris.imp.mi.fit.pooled <- pool(Sepal.Length ~ Sepal.Width +
           Petal.Width + Petal.Length + Species, data=iris.imp.mi)
display(iris.imp.mi.fit.pooled)
iris.imp.mi.fit.pooled.tab <- data.frame(
           "Estimate"=iris.imp.mi.fit.pooled@coefficients,
           "Std. Error"=iris.imp.mi.fit.pooled@ses,
           check.names=FALSE)
iris.imp.mi.fit.pooled.tab
```

```
# compare as ratio
iris.lm.res[,c(1,2)]/iris.imp.mi.fit.pooled.tab
iris.lm.NA.res[,c(1,2)]/iris.imp.mi.fit.pooled.tab
```

¿Cuál es la comparación con el conjunto de datos original, al que no le falta ningún dato? Estrictamente hablando, el paquete brms de R tendría que compararse consigo mismo y no con `lm()`, ya que por naturaleza (no sólo) con muestras más pequeñas las estimaciones difieren de forma estadística clásica y bayesiana. Lo mismo se aplica a `mi`. Nos abstenemos de hacerlo, pero téngalo en cuenta en caso de que se produzcan diferencias drásticas contrarias a lo esperado.

```
# deviances
# compare mice with amelia and brms | perspective: mice, brms
# estimates
ests <- cbind(original=iris.lm.res[,1],
              amelia=as.numeric(iris.imp.amelia.pooled$q.mi),
              mice=summary(iris.imp.mice.pooled)[,1],
              brms=fe.iris.imp.brms[,1],
              mi=iris.imp.mi.fit.pooled@coefficients,
              NA.rem=iris.lm.NA.res[,1]
)
ses <- cbind(original=iris.lm.res[,2],
              amelia.SE=as.numeric(iris.imp.amelia.pooled$se.mi),
              mice.SE=summary(iris.imp.mice.pooled)[,2],
              brms.SE=fe.iris.imp.brms[,2],
              mi.SE=iris.imp.mi.fit.pooled@ses,
              NA.SE.rem=iris.lm.NA.res[,2]
)
```

Las desviaciones vienen dadas por la relación 1-(procedimiento original/imputación))*100 en porcentaje. En primer lugar, las estimaciones de los β-pesos

```
# simple percent over-/underestimation
# for each variable over each method
apply(ests,2,function(x) 1 - iris.lm.res[,1]/x)*100
# compare methods over variables
# mean
apply(apply(ests,2,function(x) 1 - iris.lm.res[,1]/x),2,mean)*100
# median
apply(apply(ests,2,function(x) 1 - iris.lm.res[,1]/x),2,median)*100
# sd
apply(apply(ests,2,function(x) 1 - iris.lm.res[,1]/x),2,sd)*100
```

y luego las estimaciones de los errores estándar

```
# simple percent over-/underestimation
# for each variable over each method
apply(ses,2,function(x) 1 - iris.lm.res[,2]/x)*100
# compare methods over variables
# mean
apply(apply(ses,2,function(x) 1 - iris.lm.res[,2]/x),2,mean)*100
# median
apply(apply(ses,2,function(x) 1 - iris.lm.res[,2]/x),2,median)*100
# sd
apply(apply(ses,2,function(x) 1 - iris.lm.res[,2]/x),2,sd)*100
```

En cuanto a los β-pesos, los resultados difieren muy poco por término medio entre `mice` y `brms`, lo que no es sorprendente; les siguen `mi` y `Amelia`. Los errores estándar son similares, salvo que `Amelia` muestra la menor desviación con respecto al original, seguida de brms, mice y mi. La eliminación de los datos, por otra parte, resulta problemática para las estimaciones de los coeficientes. Hay una gran desviación del original para la estimación de `Petal.Width`. Para los errores estándar, la supresión de los datos tiene un efecto mucho menos devastador y las desviaciones se sitúan entre `mice` y `brms`. Como recordatorio, los NA se generaron a lo largo de MCAR.

Si no se toma la desviación media promedio sobre todas las estimaciones, sino la mediana, el orden para los β-pesos sólo cambia entre `mice` y `brms`, de lo contrario no. Para los errores estándar, nada cambia si se utiliza la Mediana como referencia en lugar de la media. Para eliminar los datos, la mediana muestra naturalmente una desviación más baja para la estimación de los β-pesos. El valor atípico, la estimación del coeficiente para `Petal.Width`, no se ve afectado. Para los errores estándar, la supresión de los datos funciona peor que `Amelia, mice` y `brms` cuando se toma la mediana, pero mejor que `mi`. La comparación de la estimación del modelo con y sin datos faltantes se hace aún más evidente cuando se explora la salida de `lm()` del conjunto de datos original frente al conjunto de datos reducido por valores faltantes.

```
display(iris.lm)
display(iris.lm.NA)
```

En el caso de Petal.Width, a diferencia de las demás variables, el coeficiente cambia en varios órdenes de magnitud

```
> -0.32/-0.04 # beta ratio
[1] 8
> 0.15/0.19 # se ratio
[1] 0.7894737
> -0.32/0.15 # t-value without Nas
[1] -2.133333
> -0.04/0.19 # t-value with NAs
[1] -0.2105263
```

Mientras que para el error estándar esto equivale a una relación de RSE = 0.789, para el coeficiente es R = 8. Esto significa que hay una diferencia de algo menos del 21% (error estándar, absoluto) frente al 700% (coeficiente, absoluto). En otras palabras, el error estándar de Petal.Width con -0.04/0.19 = -0.21 es varias veces mayor que su estimación – casi 5x – y, por tanto, este parámetro es sobre todo de estimación incierta y relativamente inutilizable. El original, en cambio, tiene un cociente de -0.32/0.15 = -2.13 y, por tanto, una estimación bastante precisa. En términos estadísticos clásicos, esto supone una diferencia en el valor *t* y, por consiguiente, en el valor *p*, que cambia mucho más allá del "umbral de significación" y sería ignorado por las personas que sólo se interesan por la significación, una decisión equivocada. Significación nos interesa menos, pero la precisión de las estimaciones es importante: si los errores estándar son mayores que las estimaciones, hay algún tipo de problema. Así que los datos que faltan, incluso si un procedimiento MCAR los produce, pueden dar lugar a estimaciones desagradables. Los procedimientos de imputación reducen el problema estabilizando las estimaciones globales y añadiendo ligeros sesgos debidos a la incertidumbre especialmente introducida, pero éstos parecen estar dentro de límites aceptables en el contexto global, como puede verse en el ejemplo.

En el presente script de R, se utilizó un valor fijo para la selección de los datos faltantes para el generador aleatorio.

> Tarea 4.11: Sustitución de valores omitidos
>
> Los lectores interesados deben repetir el ejemplo, utilizando diferentes valores para el generador aleatorio. Lo más probable es que se obtengan resultados, que pueden parecer mejores o peores. Discuta las diferencias y si son sustanciales en comparación con el conjunto de datos original.

Es importante señalar que cuando MCAR es eficaz con tamaños de muestra cada vez mayores, las estimaciones son correctas porque los datos que faltan están sujetos a un mecanismo aleatorio completo. Con muestras pequeñas, como demuestran claramente los ejemplos aquí expuestos, éste no es necesariamente el caso. Entonces, a pesar de la aleatoriedad, las estimaciones pueden desviarse fuertemente y conclusiones se basan en ellas. Sólo la replicación o el uso de un procedimiento de imputación pueden ayudar en este

caso, precisamente porque los efectos pequeños y las muestras pequeñas suelen tener un efecto en los estudios de ciencias sociales. En la práctica, sin embargo, se piensa en el sesgo de selección en los estudios clínicos, el azar puro es probablemente menos el caso estándar, sino más bien una excepción.

Las razones por las que faltan datos, la gente no responde, no participa, abandona un estudio a largo plazo, etc. pueden ser múltiples. Esto se suma a las demás influencias que afectan a los estudios. En nuestra opinión, esto por sí solo es suficiente para justificar el uso de procedimientos de imputación, por ejemplo, la imputación múltiple, y un diagnóstico exhaustivo de cómo se producen los datos que faltan en casos individuales. Para estar seguros, los análisis necesarios pueden y deben llevarse a cabo en paralelo sin y con diferentes procedimientos de imputación, examinándose la plausibilidad de los efectos y la estabilidad de los resultados entre los procedimientos, a fin de poder evaluar adecuadamente la seriedad de las conclusiones. conclusiones adecuadamente.

Entonces, ¿cuál es el mejor procedimiento de imputación? Tras esta prueba, no es posible dar una respuesta general. Sin embargo, se da el caso de que todos los procedimientos de imputación hicieron realmente un trabajo decente. La imputación múltiple mediante `mice` y el ajuste predictivo de medias merecen la pena, al igual que el enfoque bayesiano con `mi` o la imputación múltiple con `mice` o la imputación múltiple basada en el algoritmo bootstrap EM con `Amelia`. La comparación establecida entre métodos que supuestamente conducen al mismo objetivo apunta directamente al siguiente punto – procedimientos equivalentes.

4.4.9 Procedimientos equivalentes

Cuando se trata de métodos y procedimientos equivalentes, es decir, similares, cabe distinguir la equivalencia de los procedimientos y métodos de medición de la de los procedimientos y métodos de análisis de datos. Empecemos por los primeros y pasemos después a los segundos. El tema sólo se tocará ligeramente en los esquemas de cada caso, ya que es extremadamente extenso y, como se demostrará, en principio podrían utilizarse todos los procedimientos de análisis.

4.4.9.1 Equivalencia de procedimientos y métodos de medición

Si se comparan procedimientos de medición, deben recoger el mismo contenido y conducir estadísticamente a las mismas conclusiones si se utilizan los mismos procedimientos analíticos, es decir, si se aborda una cuestión equivalente. Como antecedente, por tanto, el clásico artículo de Campbell y Fiske (1959) sobre "multitrait and multimethod" puede citarse como antecedente, que es igual de relevante para los métodos mixtos en cuanto a su idea básica. Este artículo trata de la validez discriminante y convergente, y eso da en el clavo: ¿dónde empieza y acaba la convergencia o discriminancia entre métodos, procedimientos, etc.? ¿Qué es lo mismo, qué se dirige a cosas diferentes? ¿Cómo podemos comparar si sólo tenemos lo que tenemos y un criterio de verdad relativo? Si ahora tenemos datos de dos procedimientos de medición supuestamente equivalentes, podemos utilizar la comparación gráfica exploratoria (véase también EDA en el capítulo 5) antes de proceder a otras investigaciones estadísticas. Una variante gráfica sencilla y muy común es el diagrama de Bland-Altman (Bland & Altman, 1986). Este gráfico se utiliza en variaciones de estadística médica, química analítica o para datos genómicos (el denominado gráfico MA en los paquetes `Raffy, limma, marray` y `edgeR`, así como el gráfico RA basado en números enteros en `caroline`). El gráfico Bland-Altman es idéntico al gráfico (M)ean-(D)ifference de John W. Tukey (`tmd()` en `lattice`), una adaptación del gráfico cuantil-cuantil, por lo que los créditos pertenecen en realidad a Tukey. Por lo tanto, hablamos de un gráfico MD y siempre nos referimos al mismo tiempo al gráfico Bland-Altman. El gráfico MD representa la diferenzia X_1 y X2 frente a la media de las diferencias $(X_1\text{-}X_2) = 2$. Así pues, el gráfico supone una relación lineal entre los dos métodos de medición. Una relación perfecta entre dos instrumentos de medición daría como resultado una línea recta en el ángulo de 45 en el diagrama de dispersión, que pasa exactamente por el punto cero. El diagrama de Bland-Altman sólo produce ceros, es decir, una línea recta horizontal en cero. Por tanto, todas las desviaciones de cero representan una equivalencia reducida. Ahora tenemos que

preguntarnos: ¿cuánta desviación pueden tolerar los métodos de medición entre sí para seguir considerándose equivalentes? El gráfico MD comprueba

- si la variación entre dos conjuntos de mediciones es constante y
- si la varianza de las diferencias es constante en el intervalo de valores observados.

La idea es muy similar a la interpretación de un gráfico de *fitted vs. residuals* (regresión lineal) y permite extraer conclusiones diferentes:

- Una diferencia constante representa un sesgo constante y, en consecuencia, legitima el cálculo de límites equivalentes (confidencias) con respecto a la concordancia de los procedimientos o métodos que se comparan.
- Una diferencia no constante en el intervalo de valores observable requiere un diagnóstico adicional, por ejemplo mediante una regresión lineal. Además de un nuevo análisis del contenido, puede utilizarse un método como predictor y el otro como variable dependiente para aclarar la relación real.
- Si la varianza de las diferencias no es constante, puede ser necesaria una transformación adecuada de los datos para establecer una relación lineal entre los métodos y mantener la varianza constante.

El gráfico MD es fácil de implementar en R (paquetes R `BlandAltmanLeh` y `blandr`). Sin embargo, utilizamos nuestra propia función `MD.plot()`.

> Tarea 4.12: Diagrama de diferencia de medias
>
> Como tarea, los lectores interesados pueden generar vectores aleatorios con diferentes medias poblacionales o desviaciones estándar (cada una de la distribución normal, ...). Estas diferencias pueden entonces investigarse con `MD.plot()` - o pueden ser números (propios) o datos empíricos.

El siguiente código R reproduce los valores originales de Bland y Altman (1986) del paquete R `BlandAltmanLeh`, contenido en el conjunto de datos `[bland.altman.PEFR]`. Se trata de datos de flujo espiratorio (picos) de n = 17 miembros de la familia Bland, recogidos con diferentes métodos de medición y cada uno dos veces. `MD.plot()` acepta dos vectores y, además de un diagrama de dispersión de los dos vectores, produce un histograma con estimación de densidad de las diferencias entre los vectores y el diagrama MD o MA siguiendo el trabajo original de Bland y Altman (1986). El gráfico MA es una variante logaritmizada del gráfico MD. A esto se añaden los valores de confianza calculados y los cálculos intermedios (`ptII_quan_classicstats_equivalentmethods.r`). La figura 4.53 contiene los diferentes gráficos. A continuación se omiten las salidas extensas.

```
#first create data
seed <- 0987
set.seed(seed)
n <- 100
mu <- 10
sigma <- 2
x <- rnorm(n=n, mean=mu, sd=sigma)
y <- x + rnorm(n=n, mean=0, sd=1)
# mean-difference plot (Tukey)
# also known as Bland-Altman plot
res <- MD.plot(x,y)
```

Ahora una comparación con las funciones de `BlandAltmanLeh`.

```
bland.altman.stats(x,y, two=2) R-Code
```

```
bland.altman.plot(x,y, two=2, silent=TRUE, conf.int=.95)
```

El gráfico de diferencia de medias de Tukey se encuentra en el paquete R `lattice` en la función R `tmd()`.

```
# Tukey Mean-Difference Plot R-Code
tmd(xyplot(x ~ y))
```

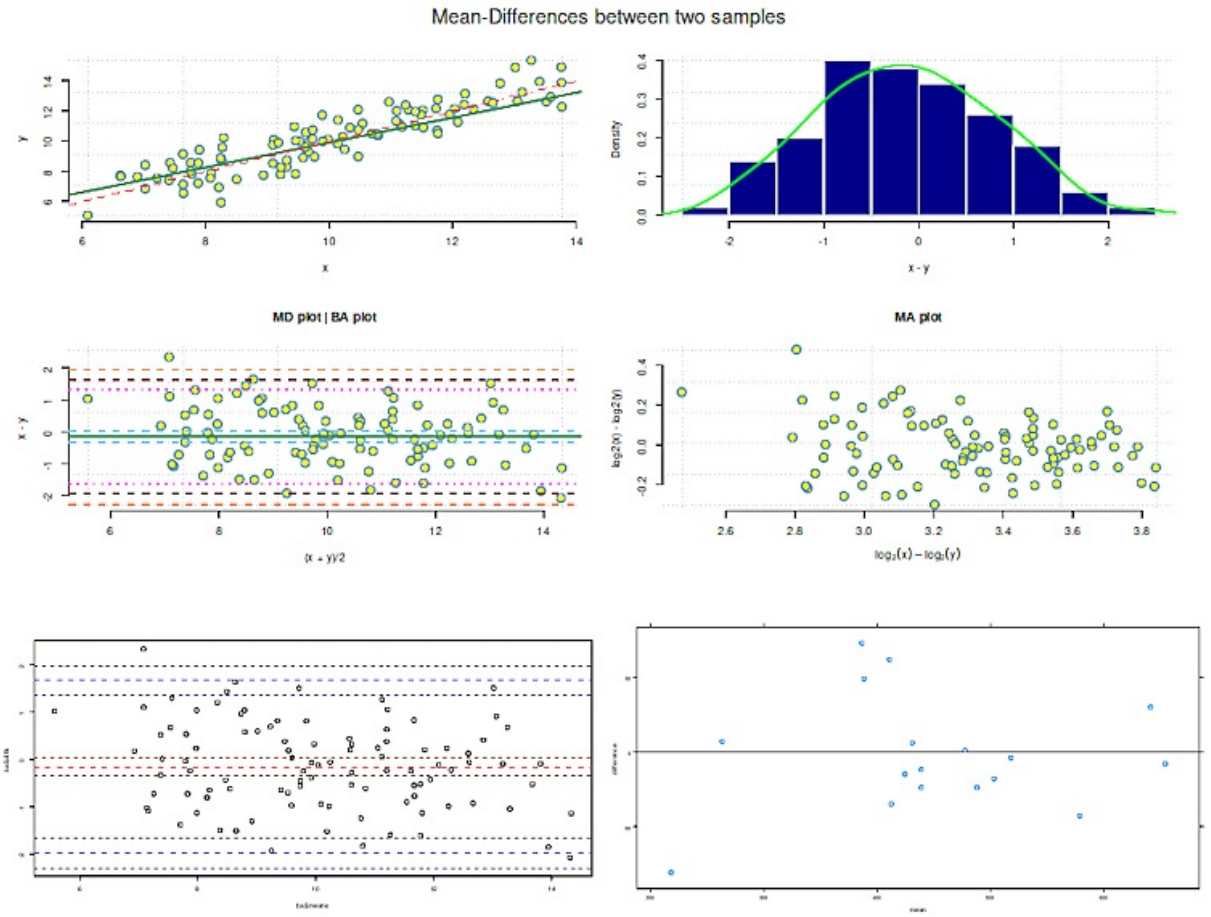


Figura 4.53. *Procedimientos equivalentes (diferencias medias, gráficos diferentes).*

Por cierto, no suele recomendarse un enfoque de correlación simple como método estándar para comparar métodos, ya que en este caso los valores atípicos pueden dar lugar a distorsiones y conclusiones falsas en función del método de medición (por ejemplo, correlaciones sobreestimadas), aunque la correlación real sea de otra naturaleza. Como de costumbre, es necesario separar cuál es la correlación real entre dos métodos (constructos) frente a la correlación basada únicamente en datos empíricos. Del mismo modo, está claro que la suposición de una relación lineal puede resultar problemática y la ausencia de una relación lineal no significa en general que dos métodos de medición no investiguen adecuadamente lo mismo. A encontrar una transformación de datos adecuada puede resultar difícil y complicado. Si se dispone de datos de replicación en lugar de datos de dos métodos de medición diferentes, puede utilizarse, por supuesto, el diagrama de Bland-Altman y los estadísticos asociados, los intervalos de confianza clásicos, es decir, los límites inferior y superior en expresiones diferentes. El procedimiento no cambia. Por supuesto, podría realizarse una simple prueba *t* entre dos procedimientos de medición o sus diferencias. Sin embargo, esto "adolece" del hecho de que en estadística clásica la hipótesis nula no puede aceptarse, sino sólo rechazarse. El no-rechazo de la hipótesis nula no supone una ganancia de conocimiento especialmente grande, porque sólo dice que el modelo utilizado no era suficiente para rechazar la hipótesis nula; pero no dice nada sobre si la hipótesis nula es cierta y en qué medida. En la estadística de Neyman-Pearson, el no-rechazo de la hipótesis nula puede llevar a un comportamiento inductivo, pero incluso en ese caso el nivel de cono-

cimiento no es alto; después de todo, Neyman-Pearson trata de decisiones, no de conocimiento. Sin embargo, establecer la equivalencia consiste precisamente en establecer la validez de la hipótesis nula o estimar la incertidumbre de que existan diferencias y en qué medida, que sean sistemáticas y no de naturaleza aleatoria. Piénsese en la introducción de un nuevo método terapéutico o de un nuevo fármaco y en la prueba de que éstos, al menos, no son inferiores a los anteriores enfoques y sustancias y, especialmente en el caso de los productos farmacéuticos, no provocan mayores efectos secundarios. Según Walker y Nowacki (2011, p.194), las comparaciones no deben hacerse con la prueba *t* tradicional por dos razones:

> „First, the burden of the proof is on the wrong hypothesis, i.e., that of a difference. In this setting, a significant result establishes a difference, whereas a nonsignificant result implies only that equivalency (or equality) cannot be ruled out. Consequently, the risk of incorrectly concluding equivalence can be very high. The other reason is that the margin of equivalence is not considered, and thus the concept of equivalence is not well defined."

Por ello, en la estadística clásica se han desarrollado las pruebas de equivalencia (Schuirmann, 1987; Walker & Nowacki, 2011) para sortear este problema básico. Según los autores (ibíd., p.193, Tabla 1), las formas de hipótesis impresas en la Tabla 4.11 (traducidas por los autores) difieren.

Tabla 4.11. *Prueba de equivalencia (Tipos de hipótesis)*

Tipo de estudio	Hipótesis nula	Pregunta de investigación
Comparación (tradicional)	No hay diferencias entre las unidades de investigación (Modo de terapía, método, etc.)	Hay una diferencia entre las unidades de investigación.
Equivalencia	Las unidades de investigación no son equivalentes.	El "nuevo" método es equivalente a los métodos anteriores.
No-Inferioridad	El "nuevo" método es inferior a los métodos (estándar) anteriores.	El "nuevo" método no es inferior a los métodos (estándar) anteriores.

El procedimiento TOST es el más utilizado. Significa "(T)wo (O)ne-(S)ided (T)ests" y caracteriza bastante bien el procedimiento. En el procedimiento TOST, se especifica un límite mínimo de equivalencia por encima y por debajo de la hipótesis nula que se va a probar, basado en el menor tamaño del efecto de interés, por ejemplo, d = 0,35. A partir de ahí, se formulan dos hipótesis nulas unidireccionales, una que viene de arriba y otra que viene de abajo, que se prueban contra la hipótesis nula de interés. Si ambas pruebas se rechazan al nivel de convención habitual, se asume la equivalencia de las pruebas, procedimientos, métodos, etc., o que la fuerza del efecto E observable empíricamente es menor que la señalada como mínimamente interesante y, por tanto, no hay efectos E significativos. No se presupone la identidad de los métodos, por lo que el término equivalencia se define como la igualdad dentro de ciertos límites de tolerancia. El método TOST se ha ampliado en el sentido de que está disponible adicionalmente para repeticiones de medición y múltiples variables en paralelo (Rose, Mathew, Coss, Lohr y Omland, 2018). El paquete TOSTER de R genera además análisis de potencia. La conexión entre la prueba de equivalencia y la prueba t normal es que la prueba t puede convertirse en principio en la prueba de equivalencia ajustando el tamaño de muestra necesario N o el nivel de tasa de error α (Siebert y Ellenberger, 2019). Si esto se omite, según los autores, la prueba t no ajustada (prueba t revisada) conduce a conclusiones sesgadas. Técnicamente hablando:

$$\alpha_{\text{Prueba de equivalencia}} = \beta_{\text{Prueba t}} \quad \text{Tasa de error tipo I} \quad (4.38)$$

$$\alpha_{\text{Prueba t}} = \beta_{\text{Prueba de equivalencia}} \quad \text{Tasa de error tipo II} \quad (4.39)$$

Wellek (2010) ofrece una visión general de las pruebas de equivalencia en función del área del problema. Se pueden encontrar en R en los paquetes TOSTER, equivalence y equivalenceTest. Lakens, Scheel e Isager (2018) muestran cómo utilizar las funciones de R en TOSTER como parte de un artículo sobre pruebas de equivalencia en la investigación psicológica. Más detallada respecto al código R es la viñeta (Lakens, 2018-08-03) sobre TOSTER. En otro artículo, Lakens (2017a) analiza las posibilidades de las pruebas de equivalencia aparte de las comparaciones de dos grupos, por ejemplo, para correlaciones o metaanálisis. Robinson, Duursma y Marshall (2005) abordan el uso de TOST en la regresión lineal y aplican sus sugerencias en el paquete R equivalence. En el sitio web de Magnusson (2016) se puede encontrar una aplicación web interactiva sobre el tema.

En primer lugar, aplicamos la prueba *t* simple para muestras dependientes a los datos de Bland-Altman- (ptII_quan_classicstats_equivalentmethods.r),

```
> # simple t-test
> # Student
> t.test(x,y, alternative="two.sided", paired=TRUE, var.equal=TRUE)

Paired t-test
data: x and y
t = -1.524, df = 99, p-value = 0.1307
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.32077174 0.04207938
sample estimates:
mean of the differences
-0.1393462

> # Welch
> t.test(x,y, alternative="two.sided", paired=TRUE, var.equal=FALSE)

Paired t-test
data: x and y
t = -1.524, df = 99, p-value = 0.1307
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.32077174 0.04207938
sample estimates:
mean of the differences
-0.1393462

> # Cohens d
> cohensd(x,y)
d|mean sd d|pooled sd
0.06807645 0.06807645
```

y añadimos el análisis con el paquete TOSTER de R. Para ello seleccionamos un efecto pequeño según Cohen con d = 0.2 y un nivel de tasa de error tipo I convencional del 5%. Dejamos la potencia para posteriores análisis de potencia en el 80% habitual.

```
> # x y dataset
> head(xy <- data.frame(x,y))
    x          y
1  9.885973 11.712461
2 10.539990 11.150191
3 11.748943 10.691732
4  9.619307 10.563684
5 12.201396 13.126682
6  8.212482  6.913318
> N <- dim(xy)[1]
> paur <- 0.8
> alpha <- 0.05
> # small d
> d <- 0.2
```

```
> dataTOSTpaired(data=xy, pairs=list(c(i1="x",i2="y")),
+          low_eqbound=-d, high_eqbound=d, alpha=alpha,
+          desc=TRUE, plots=TRUE)
TOST PAIRED SAMPLES T-TEST
TOST Results
                    t      df  p
x y     t-test -1.524003  99 0.1306962
    TOST Upper -3.524003  99 0.0003226
    TOST Lower  0.4759972 99 0.3175626

Equivalence Bounds
                Low    High    Lower   Upper
x y  Cohen's d -0.2000 0.2000
           Raw -0.1829 0.1829 -0.2912  0.01247

Descriptives
  N   Mean      Median     SD       SE
x 100 10.03538  9.824267  1.951255 0.1951255
y 100 10.17472 10.10893   2.138285 0.2138285
```

La figura 4.54 muestra la salida gráfica de `dataTOSTpaired()` para d = 0.2 (arriba) y d = 0.7 (abajo).

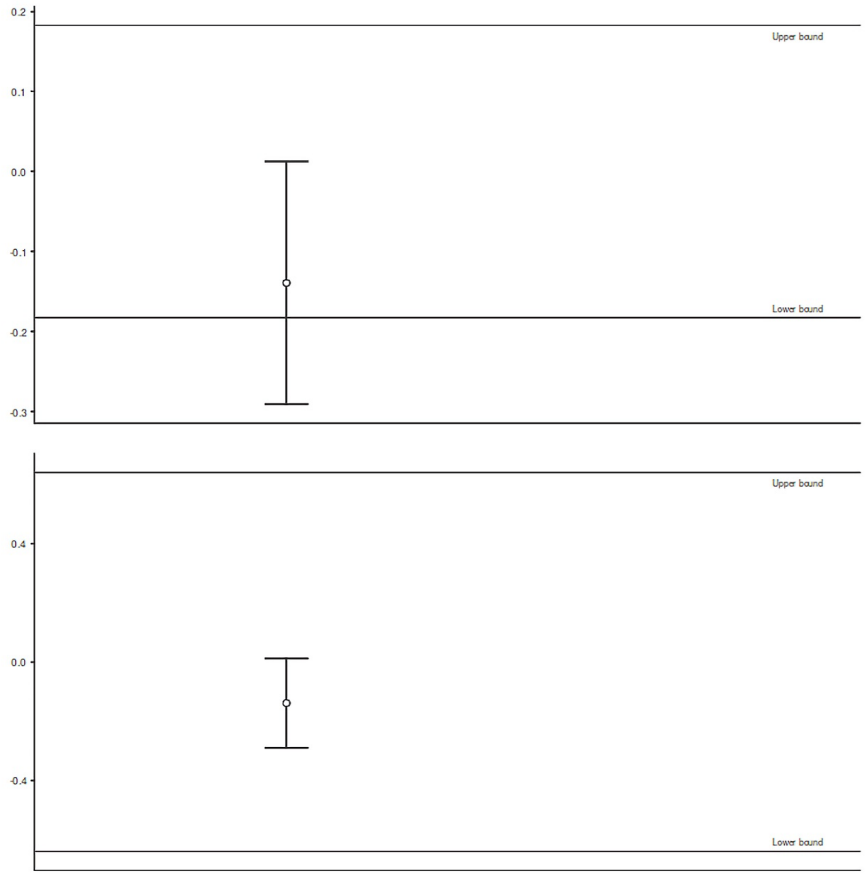


Figura 4.54. *Prueba t del TOST (muestras pareadas)*

Los resultados del procedimiento TOST no indican diferencias significativas entre los grupos para los límites superior o inferior, suponiendo un nivel α convencional del 5%. Pero no debe considerarse ningún efecto sin un análisis de potencia. La función de R `powerTOSTpaired()` calcula el tamaño de muestra necesario, la potencia alcanzada o el tamaño del efecto asumiendo un tamaño del efecto verdadero de cero. Al llamar a la función deje una variable vacía, que se calculará a partir de todas las demás variables dadas.

Partiendo de una potencia supuesta de $p = 0.8$, los límites del tamaño del efecto según Cohen de $d = 0.2$ y un nivel de tasa de error habitual, el tamaño de la muestra necesario da como resultado a

```
> powerTOSTpaired(alpha=alpha, statistical_power=paur,
            low_eqbound_dz=-d, high_eqbound_dz=d)
The required sample size to achieve 80 % power with equivalence
bounds of -0.2 and 0.2 is 215 pairs
[1] 214.0962
```

Evidentemente, la muestra actual es completamente insuficiente, con N = 17, para un efecto verdadero efecto de cero dentro de los límites de equivalencia superior o inferior de un efecto pequeño según según Cohen. Por lo tanto, no debería salir nada más para los cálculos de potencia y tamaño del efecto. La potencia post-hoc se comprueba con el tamaño de muestra dado de N = 17,

```
> powerTOSTpaired(alpha=alpha, N=N,
low_eqbound_dz=-d, high_eqbound_dz=d)
The statistical power is 27.78 % for equivalence bounds of -0.2 and 0.2 .
[1] 0.2777876
```

y no es sorprendente que no se dé potencia aquí, con los mismos parámetros que arriba. Por último, está la cuestión de los límites equivalentes (potencia del efecto según Cohen) con parámetros por lo demás constantes.

```
> powerTOSTpaired(alpha=alpha, statistical_power=paur, N=N)
The equivalence bounds to achieve 80 % power with N = 100
are -0.29 and 0.29 .
[1] -0.2926405 0.2926405
```

Así pues, sería necesario un efecto significativamente mayor para encontrar un efecto en el contexto de los demás parámetros. ¿Son los grupos equivalentes entre sí o no? Según el procedimiento TOST, lo son, pero según el análisis de potencia, la potencia dista mucho de ser suficiente para encontrar lo contrario de un efecto de cero según las características disponibles de la muestra.

Tabla 4.12: *El problema Behrens-Fisher (Hypótesis)*

Promedio \ Distr. estándar	igual	desigual
igual	H_1	H_2
desigual	H_3	H_4

¿Qué más destaca? Según el problema de Fisher-Behrens (Bretthorst, 1993 para explicaciones y una solución bayesiana analíticamente exacta), es decir, la cuestión de valores medios o desviaciones estándar desiguales/iguales, la prueba *t sólo* examina las diferencias de localización, pero no las diferencias de varianza. Sólo distingue si las varianzas se consideran iguales o desiguales, pero no prueba si realmente es así. Por lo tanto, la equivalencia sólo se comprueba de forma inadecuada con la prueba *t* o el procedimiento TOST, porque la equivalencia puede existir o no a nivel de los valores medios, pero una comparación de las varianzas conduce a un resultado completamente diferente. Mediante la estadística de Bayes, se da respuesta al problema completo de Behrens-Fisher, como se resume en la Tabla 4.12 y que conduce a cuatro afirmaciones probabilísticas, o más bien a sus negaciones:

1. Las medias son iguales y las desviaciones estándar son iguales ($=H_1$)
2. Las medias son desiguales y las desviaciones estándar son iguales ($=H_2$)
3. Las medias son iguales y las desviaciones estándar son desiguales ($=H_3$)
4. Las medias son desiguales y las desviaciones estándar son desiguales ($=H_4$)

Para aclarar la homogeneidad de las varianzas (véase el capítulo 4.4.11), podría utilizarse una prueba correspondiente para varianzas correlacionadas, como la prueba de Morgan-Pitman (Morgan, 1939; Pitman, 1939), porque se pueden suponer varianzas correlacionadas para muestras dependientes. Esta prueba, sin embargo, reproduce el problema de la prueba *t* y la falta de conocimiento ante el no rechazo de una hipótesis nula. Por lo tanto, se necesitaría un procedimiento TOST para las varianzas. Aún mejor sería una prueba que tuviera en cuenta *tanto las diferencias como las varianzas* de forma robusta.

Antes de aplicar la prueba de Morgan-Pitman, merece la pena examinar la correlación de las desviaciones al cuadrado de la media muestral para ambas muestras como medida simple de la correlación de las varianzas.

```
# correlation of deviations^2 from mean
r.dev2s <- cor((x-mean(x))^2, (y-mean(y))^2)
data.frame(r=r.dev2s,R2=r.dev2s^2)
```

Los cuadrados de desviación están obviamente correlacionados entre sí. La prueba de Morgan-Pitman proporciona ahora,

```
> # test variance of correlated data
> BA.paired <- paired(x,y)
> plot(BA.paired) # not printed here
> # Pitman-Morgan Test
> # manual Pitman-Morgan Test
> # https://link.springer.com/article/10.1186/s40488-015-0030-z
> Var.test(x,y,paired=TRUE)

Paired Pitman-Morgan test

data: x and y
t = -2.1224, df = 98, p-value = 0.03632
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.7016945      0.9882006
sample estimates:
variance of x  variance of y
3.807395       4.572264
```

que las varianzas de los dos grupos no difieren significativamente a un nivel convencional como $\alpha = 0.05$ según NHST. La prueba de Morgan-Pitman puede calcularse manualmente ad-hoc en el programa R

```
> # manual Pitman-Morgan Test
> # https://link.springer.com/article/10.1186/s40488-015-0030-z
> N <- (length(x))
> r.xy <- cor((x+y),(x-y))
> r.xy
[1] -0.2096339
> tvalue <- r.xy*sqrt((N-2)/(1-r.xy^2))
> # two-sided
> pvalue <- 2*(1-pt(tvalue, df=N-2, lower.tail=T))
> alpha <- 0.04
> data.frame(r.xy, t=tvalue, p=pvalue, alpha, sig=pvalue<alpha)
    r.xy       t       p       alpha sig
1 -0.2096339 -2.12243 1.963678 0.04  FALSE
```

y llega – como debe ser – al mismo resultado. Si pasamos a la estadística de Bayes, la estimación probabilística de la validez de la hipótesis nula no supone ningún problema. Esto puede hacerse utilizando factores de Bayes (véase el capítulo 6.8.1) o la prueba exacta *t-test* según Bretthorst (1993, véase el apartado 6.15.3.8 o el estudio de caso del apartado 6.14.3) y también permite la inclusión de información previa. Kruschke (2015b, 2013b) ofrece con la ROPE (= Region of Practical Evidence / Región de Equivalencia Práctica, véase el capítulo 6.15.3.8), otro procedimiento bayesiano para determinar la equivalencia. Kruschke (2017a) señala que "[e]l TOST equivale a comprobar que el intervalo de confidencia del 90% (no

del 95%) cae dentro del ROPE. El procedimiento TOST se utiliza para decidir sobre la equivalencia con un valor de parámetro ROPE'd". Sin embargo, el autor muestra en los mismos ejemplos, que TOST no siempre es idéntico a ROPE.

TOST puede incluso aceptar un valor de parámetro mientras que NHST lo rechaza. Según el autor, el trabajo bayesiano con HDI (Highest Density Interval) y ROPE nunca conduce a tales conflictos. Según Kruschke, las conclusiones son siempre coherentes y se refiere explícitamente a Lakens (2017b, 2017c), quien asume que los factores de Bayes y el TOST son pruebas bastante similares. Un trabajo reciente de Lakens con sus colegas (Lakens, McLatchie, Isager, Scheel, & Dienes, 2018, con código R) asume no una identidad sino una alta similitud de TOST y los factores de Bayes. Una comparación de los métodos bayesianos y TOST es discutida por Hoyda, Counsell y Cribbie (2019).

El factor de Bayes de `ttestBF()` del paquete `BayesFactor` de R prueba la hipótesis nula de que la media o diferencia de medias $\mu_\Delta = 0$, es decir, la potencia del efecto normalizado $(\mu_2 - \mu_1)/\sigma$ es igual a cero. Aplicado al conjunto de datos de Bland y Altman, esto da como resultado

```
> # comparison with Bayes Factors
> boxplot(x,y)
> BA.PEFR.1 <- ttestBF(x=x, y=y, mu=0, paired=TRUE)
> BA.PEFR.1
Bayes factor analysis
--------------
[1] Alt., r=0.707 : 0.3384376 ±0%
Against denominator:
Null, mu = 0
---
Bayes factor type: BFoneSample, JZS
> BA.PEFR.1@bayesFactor
              bf         error                      time           code
Alt., r=0.707 -1.083416 4.042975e-07 Sat Jul 17 17:46:43 2021 2d91d7201b9c5
> extractBF(BA.PEFR.1)[,"bf"]
[1] 0.3384376
> bf.alt <- extractBF(BA.PEFR.1)[,"bf"]
> # OR BF.null versus BF.alt
> (1-bf.alt)/bf.alt
[1] 1.954754
```

La figura 4.55 muestra un boxplot inicial de los dos vectores de datos. Los datos son 0,257 veces más probables bajo la hipótesis alternativa de que existen diferencias sustanciales entre los grupos. Según Jeffreys (1939/1961) los factores de Bayes de $1/3 = 3 < BF < 3$ no se toman en serio (dependiendo de si la perspectiva es la hipótesis nula o la alternativa). En el caso de la hipótesis alternativa, se aplica el intervalo $1 = 3 < BF$ o $BF < 3$), lo que, sin embargo, puede dar lugar a problemas idénticos a los de la especificación de *p* en la estadística clásica. La hipótesis nula es, por tanto, preferida por los datos con

BF = (1 - 0.2546898)/0.2546898 = 2.926344.

Es casi tres veces más probable que la hipótesis alternativa. Ahora añadimos un intervalo de potencia del efecto para probar con $-0.2 < d < 0.2$. Esto produce dos hipótesis $-0.2 < d < 0.2$ o $!(-0.2 < d < 0.2)$, cada una contra la hipótesis nula. Una vez se comprueba si el efecto se encuentra en el intervalo de $d = \pm 0.2$ o fuera de este intervalo.

```
> # with null interval d=+/-0.2
> BA.PEFR.2 <- ttestBF(x=x, y=y, mu=0,
nullInterval=c(-0.2,0.2), paired=TRUE)
> BA.PEFR.2
Bayes factor analysis
--------------
[1] Alt., r=0.707 -0.2<d<0.2 : 1.355442 ±0%
[2] Alt., r=0.707 !(-0.2<d<0.2) : 0.1219923 ±0.01%
Against denominator:
Null, mu = 0
```

```
---
Bayes factor type: BFoneSample, JZS
```

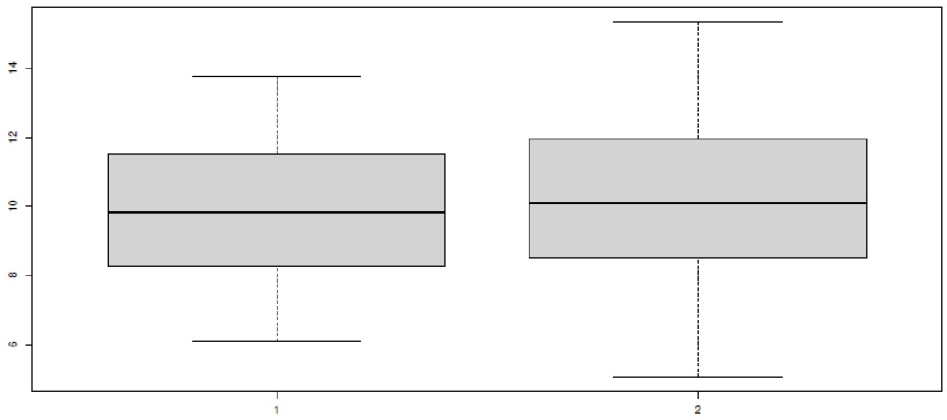


Figura 4.55. *Diferencia de medias con BayesFactor*

Las dos hipótesis pueden contrastarse entre sí, que es lo que realmente nos interesa. La prueba contra la hipótesis nula exacta es irrelevante, porque una diferencia exacta de cero es difícilmente justificable.

```
> BA.PEFR.2[1]/BA.PEFR.2[2]
Bayes factor analysis
--------------
[1] Alt., r=0.707 -0.2<d<0.2 : 11.11088 ±0.01%
Against denominator:
Alternative, r = 0.707106781186548, mu =/= 0 !(-0.2<d<0.2)
---
Bayes factor type: BFoneSample, JZS
```

La hipótesis de que las diferencias de grupo estén en el intervalo de d = $\pm$0.2 es 7,765746 veces mayor que la de que terminen fuera de este intervalo. Siempre según Jeffreys tal factor de Bayes se interpreta como una evidencia moderada. Por el contrario, la hipótesis complementaria, es decir que los datos están fuera de d = $\pm$0.2,

```
> BA.PEFR.2[2]/BA.PEFR.2[1]
Bayes factor analysis
--------------
[1] Alt., r=0.707 !(-0.2<d<0.2) : 0.09000187 ±0.01%
Against denominator:
Alternative, r = 0.707106781186548, mu =/= 0 -0.2<d<0.2
---
Bayes factor type: BFoneSample, JZS
```

da un reultado 0,1287706 veces mayor que si están dentro del intervalo especificado. Por lo tanto grupos son equivalentes entre sí dentro del rango de tolerancia especificado. Con `ttestBF()` se pueden generar valores a partir de la distribución posterior y representarlos gráficamente (véase la Fig. 4.56).

```
# draw samples from posterior R-Code
ba.PEFR.samp <- ttestBF(x=x, y=y, paired=TRUE, posterior=TRUE,
iterations=1e3)
samps <- as.data.frame(ba.PEFR.samp)
head(samps)
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l", mfrow=c(4,2))
namen <- c("substitute(mu)","substitute(sigma^2)",
"substitute(delta)","substitute(g)")
for(i in 1:4)
{
```

```
plot(samps[,i], typ="l", bty="n", col="darkred", pre.plot=grid(),
           ylab="Trace", xlab=eval(parse(text=namen[i])),cex.lab=1.2)
hist(samps[,i], prob=TRUE, pre.plot=grid(), bty="n", col="darkred",
           border="white", main="",
xlab=eval(parse(text=namen[i])), ylab="Density", cex.lab=1.2)
rug(samps[,i],col="darkorange")
lines(density(samps[,i]), col="green", lwd=2, cex.lab=1.2)
}
mtext("Samples from Posterior (t-Test with BayesFactor)", outer=TRUE,
           line=-2, cex=1.3, side=3)
```

Estos no muestran diferencias significativas con respecto a las cadenas MCMC. El enfoque ROPE descrito por Kruschke (2013a) está incluido en el paquete R BEST y utiliza simulaciones MCMC para generar los valores de las posteriors. El parámetro de transmisión `parallel=TRUE` utiliza varios núcleos de CPU.

```
> BA.best <- BESTmcmc(y1=x, y2=y, parallel=TRUE)
Waiting for parallel processing to complete...done.
> print(BA.best)
MCMC fit results for BEST analysis:
100002 simulations saved.
     mean     sd     median HDIlo  HDIup  Rhat  n.eff
mu1   10.030  0.1998 10.030 9.641  10.426 1.000 63590
mu2   10.167  0.2186 10.168 9.744  10.601 1.000 60788
nu    55.801 34.1712 47.635 8.601 124.130 1.001 28551
sigma1 1.948  0.1457  1.941 1.672   2.238 1.000 55453
sigma2 2.127  0.1601  2.118 1.831   2.453 1.000 55230
'HDIlo' and 'HDIup' are the limits of a 95% HDI credible interval.
'Rhat' is the potential scale reduction factor (at convergence, Rhat=1).
'n.eff' is a crude measure of effective sample size.
> summary(BA.best)
          mean     median  mode    HDI% HDIlo HDIup    compVal %>compVal
mu1       10.0300  10.0300 10.0403 95   9.641  10.426
mu2       10.1674  10.1676 10.1627 95   9.744  10.601
muDiff    -0.1374  -0.1376 -0.1381 95  -0.720   0.440  0       32.2
sigma1     1.9482   1.9406  1.9345 95   1.672   2.238
sigma2     2.1272   2.1179  2.0854 95   1.831   2.453
sigmaDiff -0.1790  -0.1774 -0.1732 95  -0.601   0.239  0       19.9
nu        55.8015  47.6346 32.2640 95   8.601 124.130
log10nu    1.6717   1.6779  1.7051 95   1.164   2.162
effSz     -0.0675  -0.0675 -0.0754 95  -0.351   0.215  0       32.2
```

Con un 49.7% de probabilidad posterior existe una diferencia en las medias, con un 48% una diferencia en las varianzas y con un 49.7% una diferencia en la potencia del efecto alejado de cero entre los grupos. Por lo tanto, la decisión no es inequívoca, ya que una odds ratio de las probabilidades posteriores para los tres porcentajes es aproximadamente 1.

Existen varias posibilidades gráficas para el resultado, de las que sólo mostramos las llamadas. Comencemos con la distribución posterior de las diferencias de las medias,

```
plot(BA.best)
```

y añadamos el ROPE para el rango ± 70 (= ¡escala original!) así como el valor de comparación cero y un IDH del 95%.

```
plot(BA.best, which="media", credMass=0.95, compVal=0, ROPE=c(-70,70))
```

Esto muestra que el 90% del IDH se encuentra dentro de ROPE. También se pueden comparar los parámetros entre sí, lo que resulta útil para diagnosticar la calidad del modelo.

```
pairs(BA.best)
```

Resulta más interesante probar hipótesis específicas sobre el modelo posterior. Para ello, los valores de μ_1 y μ_2 se toman de la posterior y las diferencias se forman y se almacenan en un vector.

```
# Bayesian comparisons
mean.diff <- BA.best$mu1 - BA.best$mu2
```

A continuación, para todas las diferencias potenciales de 1 a 100 en la escala original, calculamos las probabilidades posteriores para las que el criterio variable *diferencia de medias* > *criterio* es verdadero

```
out <- vector() R-Code
for(i in 1:100) out[i] <- mean(mean.diff > i/100)
```

y representarlo gráficamente. De esta manera, se puede leer muy bien qué probabilidad posterior corresponde con qué valor de comparación de la diferencia de los valores medios. Esto es mucho más interesante que probar contra un valor fijo, ya que nos da una visión general de los datos en el contexto de condiciones cambiantes.

```
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l")
plot(out, type="l", bty="n", pre.plot=grid(), col="darkred",
ylab="p(mean.diff > comp.value)",
xlab="mean differences (comp.value)")
mtext("Posterior comparison (BEST t-Test)",
outer=TRUE, line=-2, cex=1.5, side=3)
```

Si a uno sólo le interesan determinadas hipótesis, esto también puede comprobarse directamente de forma numérica.

```
> mean(mean.diff)
[1] -0.1374184
> mean(mean.diff > 0)
[1] 0.3215436
> mean(mean.diff > .1)
[1] 0.2105158
> mean(mean.diff > .7)
[1] 0.002749945
```

Por último, existe la posibilidad de trazar el radio de ROPE alrededor de cero. (véase la Fig. 4.57). Hemos ajustado los valores según el conjunto de datos, pero sin una hipótesis de fondo más detallada.

```
plotAreaInROPE(mean.diff, credMass=0.95, compVal=0.4, maxROPEradius=3)
```

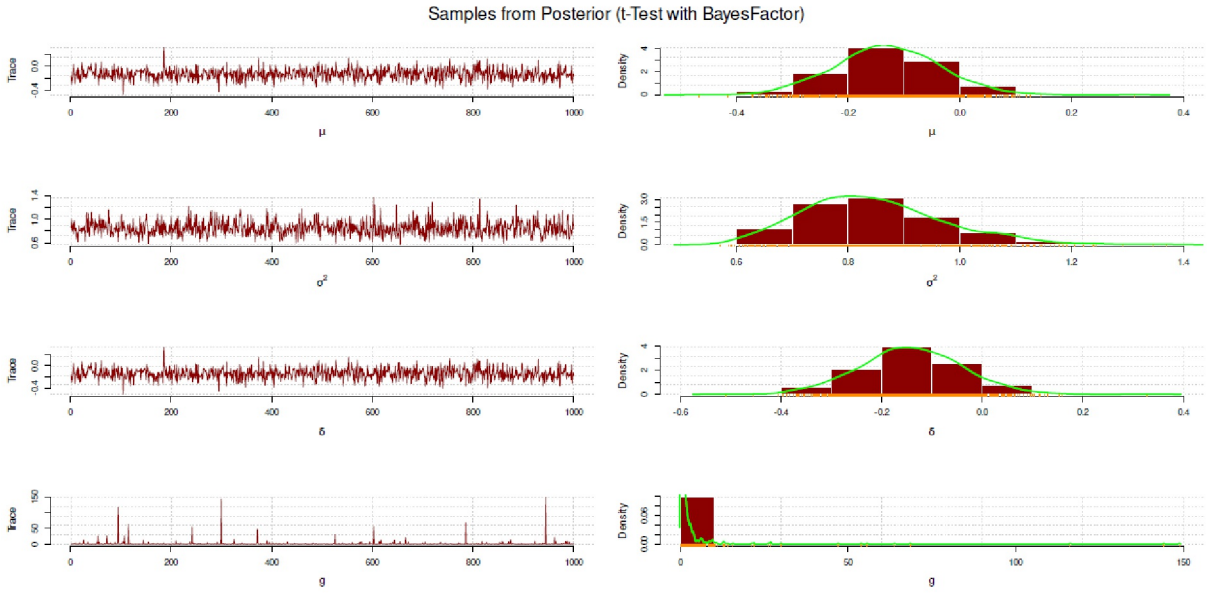


Figura 4.56. *Diferencias de media con factor Bayes (Distribuciones posteriores)*

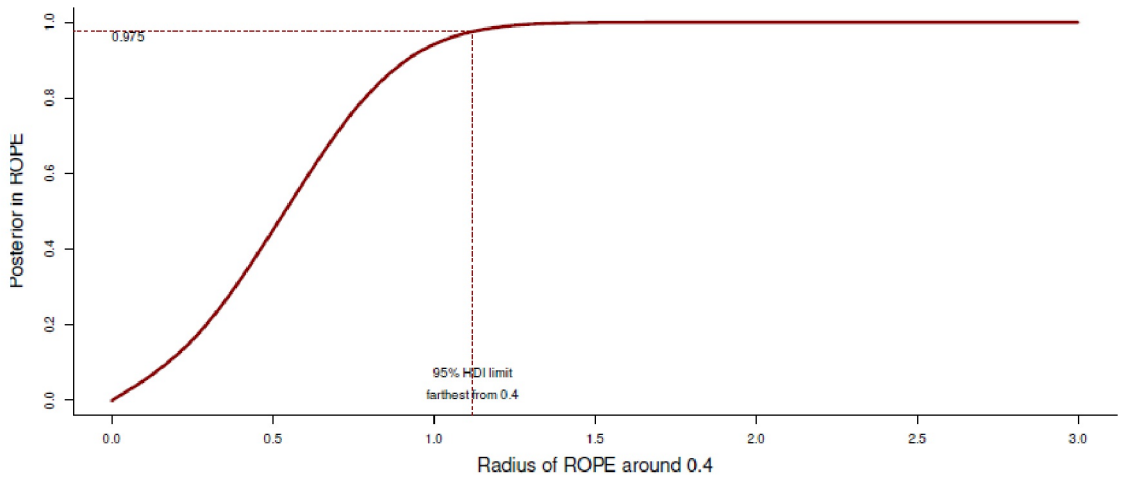


Figura 4.57. *Diferencia de media con BEST (ROPE)*

Existen otros gráficos de diagnóstico. Todos ellos son histogramas de diversos parámetros, como las diferencias de las medias, las diferencias de las desviaciones estándar, el tamaño del efecto, por grupo, etc. (véase la Fig. 4.58).

```
plotAll(BA.best, credMass=0.8, ROPEm=c(-5,5),
           ROPEeff=c(-7,7), compValm=0.4)
plotPostPred(BA.best) # only prediction, not printed
```

El último paso es la *prueba t bayesiana* exacta según Bretthorst (1993) en nuestra adaptación a R de un script `Mathematicar` de Studer (1998, s. Cap. 6.15.1 para una adaptación según el script `Mathematicar` de Gregory, 2006, `ptII_quan_classicstats_equivalentmethods.r` o `DiM_Bretthorst_UMS.r`).

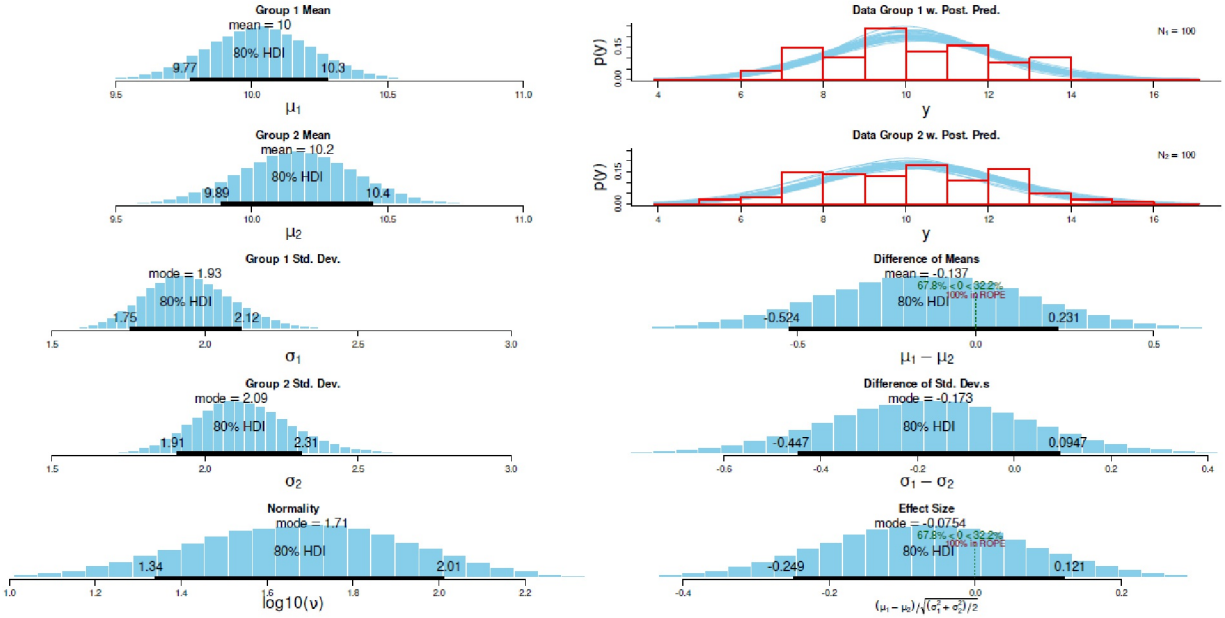


Figura 4.58. *Diferencia de media con BEST (gráfico detallado)*

También en este caso los valores iniciales previos se adaptan al conjunto de datos sin más especificaciones. Los detalles de esta solución exacta del problema de Behrens-Fisher pueden encontrarse en Bretthorst (1993) y Gregory (2006).

```
# Bretthorst (1993)
source("DiM_Bretthorst_UMS.r")
inval <- list(Si=NULL,
              Ni = N,
              Sii=NULL,
              Nii = N,
              smin = 0,
              Di = mean(x),
              si = sd(x),
              Dii = mean(y),
              sii = sd(y),
              L = 7,
              H = 14,
              sL = 2,
              sH = 6,
            snames = c("x","y")
)
DiM.BA <- DiffinMeans(inval=inval, out=FALSE)
DiM.BA$prob.df
DiM.BA$OR.df
UMSprint(resultados=DiM.BA)
```

La salida es del estilo del script original de Studer, siguiendo el artículo de Bretthorst (1993) y se reproduce íntegramente:

```
#########################################################################
###
### ON THE DIFFERENCE IN MEANS
### G.L. Bretthorst (1993)
###
### original Mathematica code by U.M. Studer (90s, Switzerland)
Note:
If any probability is printed as '1' (= one) or '0' (= zero),
it means that the probability is practically that value by giving
```

```
respect to limited computer precision.
------------------------------- Data (Input) -------------------------
N_1 = 100 : Mean_1 ± SD_1 = 10.035376 ± 1.951255
N_2 = 100 : Mean_2 ± SD_2 = 10.174722 ± 2.138285
N_total = N_1 + N_2 = 200 :
Mean_comb ± SD_comb = 10.105049 ± 2.042952
Bounds on the Mean (s_min = 0): Mean_L = 7, Mean_H = 14
Bounds on the Standard Deviation: SD_L = 2, SD_H = 6
Mean_L - Mean_comb < 0 = TRUE (-> '+'-sign between Gamma-fcts o.k.)
------------------------------- Results ------------------------------
p(mv | D_1, D_2, I) = const. 1.67957e-107
p(mbarv | D_1, D_2, I) = const. 1.99751e-108
p(mvbar | D_1, D_2, I) = const. 2.68686e-108
p(mbarvbar | D_1, D_2, I) = const. 3.25189e-109
where const. = 4.94323e-82 / p(D_1,D_2|I)
= 4.58604e+106
--------------- Model ------------------------------- Probability ---
mv: Same Mean, Same Variance: 0.77026
mbarv: Different Mean, Same Variance: 0.0916065
mvbar: Same Mean, Different Variance: 0.123221
mbarvbar: Different Mean, Different Variance: 0.0149133
----------------------------- Odds Ratios ----------------------------
The probability the means are the same is: 0.89348
The probability the means are different is: 0.10652
The odds ratio is 8.3879 to 1 in favor of the same means.
The probability the variances are the same is: 0.861866
The probability the variances are different is: 0.138134
The odds ratio is 6.2394 to 1 in favor of the same variances
The probability the data sets are the same is: 0.77026
The probability the data sets are different is: 0.22974
The odds ratio is 3.3528 to 1 in favor of the same means and variances.
------------------------------- End ----------------------------------
##########################################################################
```

Aquí la diferencia se hace evidente cuando no sólo se consideran los valores medios y las desviaciones estándar por separado, sino ambos – localización y escala – simultáneamente. Mientras que los valores de los posteriors para medias iguales o desviaciones estándar iguales son bastante idénticos al resultado de `BESTmcmc()`, la combinación de ambas hipótesis en una hipótesis global "¿son los conjuntos de datos iguales o desiguales?" muestra algo diferente (Bretthorst, 1993 o Gregory, 2006 para más detalles). En este caso, una odds ratio de 2.9509 : 1 significa que los conjuntos de datos tienen medias diferentes y/o desviaciones estándar diferentes. Desde este punto de vista, la equivalencia se da a nivel local (valores medios, desviaciones estándar), pero no a nivel global (valores medios y/o desviaciones estándar combinados). Incluso si se aumenta el intervalo de tolerancia equivalente de $440 < \mu_0 < 480$ (valores medios) o $100 < \sigma_0 < 120$ (desviaciones típicas) a $380 < \mu_0 < 520$ (valores medios), la probabilidad posterior para valores medios iguales cambia a 60.05%, pero la odds ratio posterior global para los mismos conjuntos de datos sólo cambia a 2.3212 : 1. Sigue siendo más del doble de probable que los conjuntos no tengan las mismas medias y/o desviaciones estándar que viceversa. Los lectores interesados pueden reproducir esto pasando estos nuevos valores a la variable `inval`, manteniendo todo lo demás constante, y simplemente llamando de nuevo a `DiffinMeans()`. Ahora tenemos que decidir si un cociente de probabilidades de este tipo se considera sustancialmente significativo o no. Nos gusta mantenernos alejados de las interpretaciones radicales, como es habitual con los factores de Bayes.

Aunque las soluciones de `TOST` y Bayes no se contradicen en cuanto a las conclusiones, los detalles y especialmente los gráficos muestran claramente que los métodos tratan de forma diferente la información disponible y la procesan en consecuencia. La solución de Kruschke y la *prueba t bayesiana* exacta de Bretthorst llegan a soluciones prácticamente idénticas (probabilidades posteriores o Odds Ratios posteriores), aunque Kruschke trabaja con MCMC mediante `JAGS`, mientras que Bretthorst basa su solución en el cálculo integral analítico puro. El valor *p* de TOST como $p(D|H_0)$ no puede relacionarse directamente con la probabilidad posterior de los enfoques de Kruschke y Bretthorst, que van mucho más allá de la cuestión de `TOST` y formulan afirmaciones sobre todo el problema de Behrens-Fisher. El análisis con el factor de Bayes lleva a una conclusión comparable, por la que debe tenerse en cuenta que el factor de Bayes, como cociente

de probabilidades (Likelihoods) marginales – es decir, la actualización de las hipótesis previas sobre las hipótesis basadas en los datos $p(D|H_1)/p(D|H_2)$ – no arroja ninguna probabilidad posterior que pueda decir algo sobre la probabilidad de las hipótesis a la vista de los datos dados, es decir, $p(H|D)$. En consecuencia, no se puede comparar numéricamente de modo directo con el análisis bayesiano, ya que los factores de Bayes apuntan a algo puramente cualitativo.

4.4.9.2 Equivalencia de los procedimientos y métodos de análisis de datos

Siguiendo la línea de lo dicho hasta ahora, podemos considerar no sólo la relación entre los métodos de medición, sino también la de los métodos de análisis de datos. ¿Cómo podemos elegir entre análisis similares que no son idénticos, pero que son perfectamente legítimos para analizar los mismos datos? Dado un criterio relativo de verdad, se requiere una justificación racional y adaptada al caso, basada en el conjunto de datos concretos, en lugar de una justificación última imposible pero eternamente válida. En caso de duda, realizamos ambos o incluso varios análisis y comparamos los resultados para comprender qué hace cada uno de ellos con los datos.

La cuestión de la equivalencia de los métodos de análisis de datos ya ocupa el posterior punto clave de la discusión sobre métodos mixtos en el capítulo 13. Allí surge el problema de cómo seleccionar diferentes instrumentos de investigación y procedimientos de análisis de datos de modo que se pueden integrar los resultados respectivos. Si se realiza esta integración al nivel CUAN-CUAL, CUAN-CUAN o CUAL-CUAL es, en última instancia, insignificante. En principio, procedimientos equivalentes deberían conducir a resultados equivalentes. Si no es así, hay que seguir investigando.

Sin embargo, dado que no existe un estándar de oro fijo para determinar la equivalencia se encuentra como enfoque alternativo la validación cruzada. En este caso, los datos empíricos se dividen aleatoriamente en un conjunto de entrenamiento y un conjunto de validación. Esto puede hacerse en una proporción de 80:20 o de otro modo. A continuación, los métodos analíticos se ajustan al conjunto de entrenamiento y se comprueba lo bien que predicen el conjunto de validación basándose en los datos de entrenamiento. El procedimiento podría combinarse con métodos de medición equivalentes de modo que, en el caso más sencillo, dos métodos de medición se encuentran con dos métodos de análisis. Pueden encontrarse ejemplos de un procedimiento de este tipo, por ejemplo, en concursos estadísticos, cuando el objetivo es predecir algo basado en criterios previamente especificados. Conocido es el concurso para predecir quién sobrevivió en el Titanic (estudio de caso exploratorio en el capítulo 5.5.4) utilizando procedimientos de análisis de datos de libre elección. Para ello se utilizó el conjunto de datos existente sobre el Titanic (quién sobrevivió, qué características tenían las personas, etc.) fue dividido y analizado por varios investigadores (la mayoría con R o Python) de forma que se maximizara la predicción de supervivencia. El vencedor es el modelo estadístico que mejor predice las tasas de supervivencia y diferencia al máximo entre supervivientes y no supervivientes basándose en los datos brutos.

La ventaja de la *validación cruzada* es que la elección del método de análisis de datos es independiente del método (la validación cruzada). Esto significa que, en principio, se pueden utilizar todos los métodos posibles que se cuestionan en términos de contenido. En el caso de Titanic, se podrían tomar árboles de regresión o algoritmos de cluster, siempre que predigan correctamente la supervivencia. Tal empeño puede ampliarse de modo que todo el proceso de dividir el conjunto de datos, el entrenar los algoritmos de regresión y precedir la supervivencia pueda llevarse a cabo mediante *Bootstrap* que se repita muchas veces y no sólo una vez. De los resultados se pueden deducir los errores estándar y las estimaciones de robustez y con esto hacer afirmaciones globales sobre qué método predice mejor los datos, si también se utilizan diferentes métodos de análisis de datos. Sin embargo, todo esto presupone que, en el plano de la validez de contenido, los métodos de análisis en cuestión son equivalentes en términos de contenido. Un paso más sería considerar también criterios lejos de los datos generados por los métodos de medición investigados, lo que daría lugar a afirmaciones sobre la validez externa de los instrumentos.

Para comprender mejor la lógica de decisión de la determinación de equivalencia, cabe preguntarse ¿Qué ocurre si se toma el algoritmo de prueba equivocado para calcular una variable de prueba o si los datos empíricos en relación con la población no cumplen los requisitos para el procedimiento de prueba seleccionado? En caso de que falten requisitos previos, el nivel de escala o los supuestos de distribución sean

erróneos o supuestos de distribución, podemos hablar de un error. Pero, ¿y si más de un procedimiento de análisis de datos se ajusta perfectamente a los datos y a las respectivas condiciones previas, pero conduce a diferentes resultados y, en consecuencia, conclusiones diferentes?

Si, al comparar métodos de medición, se sigue estrictamente la "lógica binaria de significación" de *uno u otro*, en cada uno de estos casos se utiliza una distribución de prueba ligeramente diferente para calcular el valor *p*, lo que, en sentido estricto, siempre es erróneo, ya que no existen modelos absolutamente verdaderos. Los valores p diferirán definitivamente, aunque sólo sea en decimales, porque un algoritmo diferente está procesando los datos. Así surgen valores *p* demasiado conservadores o demasiado permisivos, si se toman demasiado en serio los valores *p* exactos o si sólo se respetan los límites convencionales para determinar la significación estadística. Esto tiene un efecto más o menos drástico en el rechazo o la retención de las hipótesis (nulas), dependiendo de la teoría estadística que se siga en el propio trabajo. En este caso se carece de un criterio absoluto, por lo que se necesita un enfoque diferente de los datos para medir e interpretar sin ambigüedad una desviación al alza o a la baja. La fe ciega en los valores *p* no es útil en este caso. En consecuencia, sigue sin estar claro cuáles son exactamente las diferencias, ya que falta el criterio de referencia absoluto. ¿En qué resultado se "confía" ahora, sobre cuya base se decide? Más adelante en el texto se presenta un ejemplo de datos basado en varias pruebas de distribución normal. Empecemos con un ejemplo mental:

- Los métodos A y B son ambos métodos legítimos de análisis de datos para un problema en términos de contenido y
- se cumplen todas las condiciones de los procedimientos A y B
- El procedimiento A conduce a un valor p de $p = 0.045$
- El procedimiento B conduce a un valor p de $p = 0.054$
- ¿Qué conclusiones se extraen ahora? ¿Son fiables o serias?

La cuestión debería estar ahora más clara. Tal situación se centra en la necesidad de encontrar el mejor procedimiento frente a todos los demás, lo que no es necesariamente una actitud sensata. Gran parte de este libro se ha dedicado a no probar modelos entre sí, sino de formular modelos complejos que integren casos especiales y mantengan la complejidad. Para ello es necesario pasar de una lógica "o lo uno o lo otro" a una actitud "ambos/y", pero sin descuidar la precisión científica.

El uso de un límite convencional del 5% en el ejemplo anterior conduce directamente a incoherencias, a saber, cómo justificar esta elección al margen de la convención. ¿En qué pueden entonces criterios de decisión legítimos y serios? No hay que olvidar que las muestras son habitualmente aleatorias y cambiantes. En consecuencia, nos interesan las estimaciones más allá de las fluctuaciones de la muestra. Es difícil ver que sólo se puede utilizar un método a la vez – método de medición o procedimiento de análisis de datos – sea el generalmente válido y todos los demás no se encajan. Parece mucho más sensato combinar varias perspectivas para llegar a conclusiones coherentes. Si se comparan procedimientos estadísticamente, se presupone que son coherentes en cuanto al contenido o que se conoce la correlación de los constructos subyacentes. De lo contrario, las conclusiones no llevan a ninguna parte. Para el ejemplo anterior se necesita una solución creativa, y desde luego no es la única. Puede incluir

- Abandono total de los valores p y las barreras de significación.
- Concentración en el contenido.
- Siguiendo a Gelman y Hill (2007), tomarse en serio los valores $t \geq 2$, si la magnitud y la dirección de los efectos son coherentes con los supuestos teóricos (véase el análisis del diseño, capítulo 4.3.3.2). En casos concretos los valores $t \geq 2$ pueden, sin embargo, ser justificadamente relevantes.
- Utilización de procedimientos EDA (véase el capítulo 4.2.3.2) para comprender la lógica de los procedimientos de análisis de datos utilizados en cada caso y sus diferentes influencias en los datos.
- Uso selectivo de distintos procedimientos para aprender de las diferencias y los puntos en común frente a los datos y comprender el tema con mayor profundidad. No se trata entonces de excluir procedimientos, sino de utilizarlos para profundizar en la comprensión cuando los datos no estén bien explicados, hay lagunas, etc.

No se trata de que los métodos de análisis de datos puedan destilar *objetivamente* algo de los datos sin *cambiarlos*. Al contrario, en cada análisis queda claro que el procedimiento añade y presupone supuestos que no están (o no pueden estar) fundamentados en los propios datos y, por tanto, no sólo los modifican potencialmente, sino que lo hacen únicamente a través del propio análisis. Se presupone un modelo y tiene de verificarlo. Se repite aquí la figura del principio de incertidumbre de Heisenberg. No sólo es eficaz en los niveles cuánticos, sino representa también un fenómeno fundamental del mundo sensorial. La percepción es construcción, y cuanto más se centra la atención en un aspecto, lo menos recursos hay para procesar al mismo tiempo otros aspectos con la misma calidad. Si los físicos objetan ahora que los no físicos presumiblemente no entienden a Heisenberg lo suficiente como para poder aplicarlo, probablemente esto sea correcto en lo que respecta a la física. Sin embargo, la objeción no viene al caso. No se trata de física, sino de lo que podemos para nuestros campos de investigación a partir de las metáforas e ideas, conceptos e intuiciones y percepciones. Así hacemos la experiencia repetible de que en un mundo limitado y relativo que cambia constantemente, nuestra propia influencia y capacidades de percepción son limitadas. Si se aplican recursos a un objeto de una determinada manera, esto suele llevar a que otras variables bastante desconocidas pero influyentes pasen a un segundo plano. Por eso algunas cosas se investigan en el laboratorio, porque fuera de él las influencias se vuelven inmanejables. Pero incluso en el laboratorio sólo podemos investigar una cosa. Intentar varios aspectos a la vez que estén interrelacionados y tengan influencia unos sobre otros – en física es la velocidad y la ubicación, sólo podremos hacerlo de forma inadecuada. *Todo al mismo tiempo no es posible*. La cuestión de qué método es superior evoluciona entonces hacia la cuestión de cómo un método cambia exactamente *los datos y la perspectiva* que adopta sobre los datos. Esto lleva otro paso al primer plano, a saber, dejar de buscar explicaciones ausentes o erróneas por parte de los propios modelos, sino buscar las fuentes de error, o sea cambios intencionados y no intencionados que causa principalmente la aplicación de modelos a los datos.

Si ahora reducimos el campo a los procedimientos de medición y su legitimidad en la estadística clásica, se hace evidente que algunos procedimientos de prueba reaccionan de forma relativamente robusta a las violaciones de sus previas y otros no. Las razones de estas violaciones pueden residir en los propios datos (fluctuaciones aleatorias de la muestra) y, por tanto, en la materia o *simplemente* reflejar problemas de diseño y medición o una comprensión limitada de la materia, es decir, del modelo supuesto. Los posibles sesgos son pronunciados en diferentes direcciones y en diferentes grados dependiendo del método sin que exista un criterio vinculante para especificarlos en términos absolutos. Para conocer detalles concretos de este problema extremadamente complejo, hay que consultar la bibliografía pertinente sobre el procedimiento de análisis que debe utilizarse en cada caso, lo que complica la situación general a la hora de tomar decisiones concretas (por ejemplo, la elección de un procedimiento específico). Una transferencia al contexto empírico no suele ser trivial, sobre todo cuando se trata de casos límite. Los hábitos y las convenciones guían entonces la acción, por ejemplo porque determinadas preguntas se suelen investigar con más frecuencia en la bibliografía pertinente o porque una comunidad científica prefiere claramente determinadas clases de métodos. En psicología, esto se refleja en el uso de modelos de ecuaciones estructurales y en la investigación educativa empírica (por ejemplo, la medición de competencias) por la preferencia por los modelos de Rasch. Tenemos que preguntarnos críticamente si estos modelos son realmente los que prometen la máxima ganancia de conocimiento en casos individuales.

Volvamos a la estadística clásica y a los pre-requisitos típicos de los procedimientos de prueba. Estos sonla *distribución normal de los residuos*, la *homogeneidad de la varianza*, la *identificación de valores atípicos* y el tema de los *puntos de datos influyentes*.

4.4.10 Distribución normal de los residuos

Los requisitos previos de las pruebas clásicas suelen incluir el supuesto de una distribución normal de los residuos y la homogeneidad de la varianza.

Los residuos deben anularse mutuamente en muchas unidades de estudio con valor esperado = 0 y varianza denotada y aplicarse por igual a todas. Todo lo que el modelo no puede procesar son residuos (¡no necesariamente errores!) y éstos no deben basarse en un sistema o un patrón identificable, ya que la

existencia de dicho patrón significaría que el modelo estaría incompleto. Se pueden detectar las violaciones trazando los residuos frente a la distribución normal.

En primer lugar, un ejemplo de Dobson (1990, p.9) sobre el peso de las plantas (grupo de control, grupo de tratamiento) de la página de ayuda para `lm` (regresión lineal, ver Fig. 4.59, `ptII_quan_classicstats_normaldist_residuals.r`). El modelo lineal ya se calculó en el capítulo 4.3.4. Empezamos con el modelo lineal obtenido `lm.fit` y empezamos con los residuos.

```
# example from '?lm' (Annette Dobson) R-Code
# fitted model lm.fit already present
residual <- residuals(lm.fit)
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l", mfrow=c(1,2))
qqnorm(residual, panel.first=grid(), col="blue", cex=1.4, pch=21,
           bg="skyblue", main="", xlab="theoretical quantiles",
           ylab="empirical quantiles", bty="l")
qqline(residual, col="red", lwd=2, lty=1)
mtext("Comparison residuals versus normal distribution",
           outer=TRUE, line=-1.5, cex=1.5, side=3)
mtext(paste("linear model: ",deparse(formula(lm.fit)),sep=""),
           line=1.02)
```

La función de R `qqPlot()` del paquete de R `car` produce además límites de condensación clásicos para la distribución normal. Los datos también pueden compararse con un histograma con estimación de densidad (véase la Fig. 4.59):

```
# histogram
resid.dens <- density(residual)
xlim <- range(resid.dens$x)
ylim <- range(resid.dens$y)
hist(residual, xlim=xlim, ylim=ylim, panel.first=grid(), prob=TRUE,
           border="white", col="skyblue", main="", xlab="residuals",
           ylab="density", breaks="FD")
lines(resid.dens, col="steelblue", lwd=2, lty=1)
width <- round(1.5*range(residual),1)
sek <- seq(from=width[1], to=width[2], length.out=100)
lines(sek,dnorm(sek,mean(residual),sqrt(var(residual))),
           col="red", lwd=2, lty=1)
legend("topright", legend=c("theory","empirical"), lty=c(1,1),
           lwd=2, col=c("red","steelblue"), bty="n")
mtext("Histogram residuals", line=1.02)
```

R ofrece pruebas clásicas de distribución normal, como la prueba de Shapiro-Wilk `shapiro.test()`, la prueba de Kolmogorov-Smirnof `ks.test()` o la prueba de Anderson-Darling `ad.test()` del paquete `nortest`:

```
> # normality tests
> # shapiro
> shapiro.test(residual)
Shapiro-Wilk normality test
data: residual
W = 0.94744, p-value = 0.3299
> # kolmogorov-smirnof
> ks.test(residual,"pnorm",mean(residual),sqrt(var(residual)))
One-sample Kolmogorov-Smirnov test
data: residual
D = 0.1301, p-value = 0.845
alternative hypothesis: two-sided
> # anderson-darling
> ad.test(residual)
Anderson-Darling normality test
data: residual
A = 0.40742, p-value = 0.3166
```

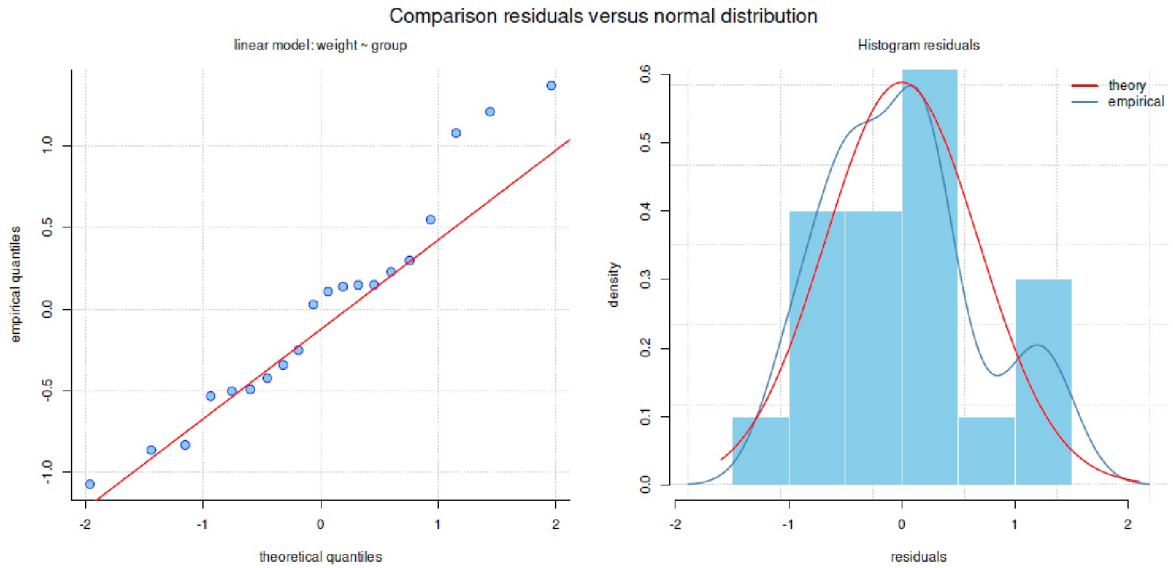


Figura 4.59. *Gráfico de residuos vs. distribución normal (diagrama de dispersión, histograma)*

Aunque todas las pruebas según la formulación de la hipótesis nula no rechazan el supuesto de una distribución normal, no aportan en modo alguno pruebas de la existencia de esta distribución – ésta es la lógica clásica de las pruebas que pueden rechazar pero no aceptar la hipótesis nula. Además se observa que los valores *p* sí difieren entre sí (Shapiro $p = 0.329$, Kolmogorov-Smirnof $p = 0.845$, Anderson-Darling $p = 0.316$) – un hecho que ya se ha discutido anteriormente con respecto a los métodos equivalentes. En el presente caso la discrepancia con la prueba de Kolmogorov-Smirnof puede explicarse por el hecho de que se sabe que es adecuada cuando se conocen las medias y las varianzas teóricas; y en caso contrario el resultado es bastante inadecuado. En el código R anterior, se han utilizado los valores de muestra empíricos medidos para simplificar las cosas, ya que no se disponía de muestras teóricas. Ciertamente no es la variante más favorable, pero corresponde al procedimiento en un contexto real.

> Tarea 4.13: Muestreo aleatorio
>
> Aunque los números aleatorios se extrajeron aquí directamente de la distribución normal teórica, no se produce necesariamente un valor p de p > 0,95 o incluso mayor. ¿Por qué? Escriba un código R correspondiente.

Para completar el debate, también podría considerarse la inclinación (skewness) y el exceso (es decir, la curvatura o la curtosis).Pero incluso en este caso, no hay reglas definitas que proporcionen un límite claro y siempre válido de la gama de valores aceptables. El resultado en la práctica lo muestra una larga discusión en una plataforma de investigación (Aslam, 2014) precisamente sobre este tema – cabe señalar que aquí participaron profesionales de la universidad y la investigación y no estudiantes o personas ajenas al campo. En resumen, esto significa que cada uno tiene su propia opinión y es capaz de justificarla, tanto en términos de contenido como con serias de referencias de libros de texto y artículos de revistas. Una respuesta generalmente válida es imposible. Por cuestión de forma recapitulemos los valores de inclinación y exceso y pasemos por alto el hecho de que la función `skewness()` del paquete R `e1071` ofrece tres fórmulas diferentes para calcular la inclinación, que proceden de libros de texto antiguos o corresponden a las definiciones en otros paquetes de software. Lo mismo ocurre con el exceso:

```
> # skewness and kurtosis, they should be around (0,3)
> # skewness
```

```
> moments:::skewness(residual)
[1] 0.4907613
> for(i in 1:3) cat(paste("type = ",i,": skewness = ",
+ round(e1071::skewness(residual,type=i),5),"\n",sep=""))
type = 1: skewness = 0.49076
type = 2: skewness = 0.53148
type = 3: skewness = 0.45442
> # kurtosis
> moments:::kurtosis(residual)
[1] 2.545242
> for(i in 1:3) cat(paste("type = ",i,": kurtosis = ",
+ round(e1071::kurtosis(residual,type=i),5),"\n",sep=""))
type = 1: kurtosis = -0.45476
type = 2: kurtosis = -0.22042
type = 3: kurtosis = -0.70292
```

Mientras que los valores de la inclinación siguen pareciendo relativamente comparables, los valores del exceso difieren mucho, pero esto se debe a las fórmulas y a la diferente interpretación en cada caso (véanse las páginas de ayuda de las funciones de R). ¿Cuál es ahora el intervalo aceptable? El paquete `normtest` de R proporciona más pruebas para la distribución normal (Jarque-Bera en dos variantes, Frosini, Geary, Hegazy-Green en dos variantes, Spiegelhalter, Weisberg-Bingham así como pruebas basadas en la inclinación y el exceso).

> Tarea 4.14: Pruebas de distribución normal
>
> El lector interesado puede leer la discusión mencionada en `researchgate.net` sobre el tema de las pruebas de distribución normal y, siguiendo la línea de la discusión mantenida aquí, ponerla en práctica en un ejemplo real. Sería bueno disponer de un ejemplo que no sea del todo obvio, sino limítrofe. De este modo, las dificultades de decisión se hacen patentes con relativa rapidez.

Y ahora, para comparar, números aleatorios de la distribución normal contra lo mismo (véase la Fig. 4.60):

```
# bigger samples R-Code
# random draws
# plot against rnorm()
seed <- 9876
set.seed(seed)
rnd.call <- c("rnorm(n=1e5, mean=100, sd=10)")
rand.nd <- eval(parse(text=rnd.call))
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l", mfrow=c(1,1))
qqnorm(rand.nd, panel.first=grid(), col="blue", cex=0.8, pch=21,
           bg="skyblue", main="", xlab="theoretical quantiles",
           ylab="empirical quantiles", bty="l")
qqline(rand.nd, col="red", lwd=1.4, lty=1)
mtext("Plot random values against normal distribution",
           outer=TRUE, line=-1.5, cex=1.5, side=3)
mtext(paste("Source: ",noquote(rnd.call),sep=""))
```

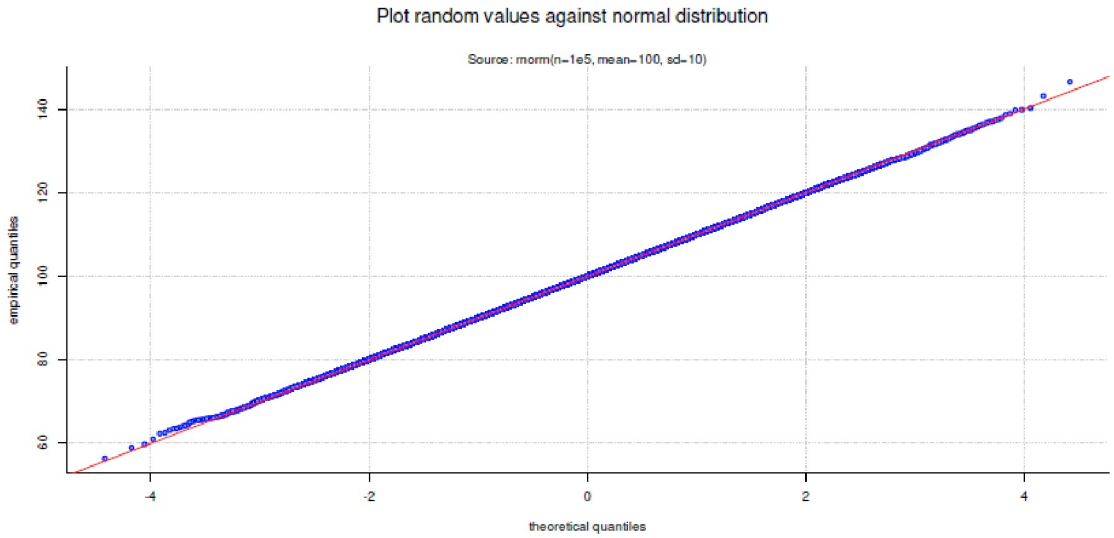


Figura 4.60. *Diagrama de dispersión (valores aleatorios de distribución normal frente a distribución normal)*

Y ahora las pruebas y valores de inclinación y exceso:

```
> # normality tests
> seed <- 9876
> set.seed(seed)
> rand.nd2 <- rnorm(n=5000, mean=100, sd=10)
> # shapiro
> # only possible sample sizes between 3 and 5000
> shapiro.test(rand.nd2)

Shapiro-Wilk normality test
data: rand.nd2
W = 0.9997, p-value = 0.6972
> # kolmogorov-smirnof
> # theoretical values possible! see above
> ks.test(rand.nd2,"pnorm",100,10)

One-sample Kolmogorov-Smirnov test
data: rand.nd2
D = 0.016343, p-value = 0.1383
alternative hypothesis: two-sided
> # anderson-Darling
> ad.test(rand.nd)

Anderson-Darling normality test
data: rand.nd
A = 0.24821, p-value = 0.7505
> # skewness and kurtosis, they should be around (0,3)
> # skewness
> moments:::skewness(rand.nd)
[1] 0.0006175403
> for(i in 1:3) cat(paste("type = ",i,": skewness = ",
+ round(e1071::skewness(rand.nd,type=i),5),"\n",sep=""))
type = 1: skewness = 0.00062
type = 2: skewness = 0.00062
type = 3: skewness = 0.00062

> # kurtosis
```

```
> moments:::kurtosis(rand.nd)
[1] 2.986258
> for(i in 1:3) cat(paste("type = ",i,": kurtosis = ",
+ round(e1071::kurtosis(rand.nd,type=i),5),"\n",sep=""))
type = 1: kurtosis = -0.01374
type = 2: kurtosis = -0.01368
type = 3: kurtosis = -0.0138
```

4.4.11 Homogeneidad de la varianza (homocedasticidad)

Por ejemplo, si se comparan grupos en el marco de un análisis de varianza, las varianzas de los grupos deben ser las mismas. En cambio, en los modelos lineales, las varianzas de los errores deben ser constantes. Las violaciones del primer caso pueden examinarse de forma aproximadamente precisa mediante histogramas y estimaciones de densidad o mediante boxplots. También es posible realizar una prueba, a saber, la prueba de Levene `leveneTest()` (Levene, 1960) del paquete R `car`. El paquete `lmtest` de R implementa la prueba de Breusch-Pagan `bptest()` (Breusch & Pagan, 1979) para la heteroscedasticidad. En el caso de modelos lineales, la varianza de error no constante puede identificarse mediante un gráfico de nivel de dispersión `spreadLevelPlot()` de los residuos (residuos estudiados) frente a los valores estimados por el modelo (Fox, 2002; véase la Fig. 4.61) y mediante `ncvTest()`, también todos de `car`. Schützenmeister, Jensen y Piepho (2012) describen un método para examinar gráficamente tanto el supuesto de distribución normal como el de homocedasticidad en los GLM. Más pragmático en el caso de varianzas de error no constantes es el uso de HLMs/ MLMs. Éstos sólo tienen ventajas sobre el análisis de varianza convencional. En particular, el uso de niveles elude elegantemente los sesgos debidos a violaciones de prerequisitos (por ejemplo, la violación de la homogeneidad de la varianza). En consecuencia, Gelman y Hill (2007) recomiendan utilizar HLM/MLM siempre que sea posible. Los estudios deben planificarse en consecuencia. Sin embargo, a la inversa, esto demuestra que la elección de la metodología ya tiene una influencia duradera en el diseño y, por tanto, en los posibles resultados. El paquete R `PairedData` ofrece otras pruebas con procedimientos robustos que calculan, entre otras cosas, las desviaciones absolutas de la mediana. Para obtener una visión general de las pruebas disponibles, escriba `library(help=PairedData)` en R. En las pruebas de parámetros de escala se incluyen Sandvik y Olsson (1982), McCulloch (1987), la prueba ampliada de Brown-Forsythe (Brown y Forsythe, 1974; Wilcox, 1989), Grambsch (1994) y Bonett y Seier (2003) y Conover e Imam (1981).

> Tarea 4.15: Interpretación de la homocedasticidad
>
> Observa la salida de R e investiga en las páginas de ayuda y en las fuentes estadísticas relevantes. fuentes estadísticas relevantes cómo interpretar los valores y gráficos anteriores. Observa también el ejemplo de la página de ayuda de ncvTest() y explica la(s) diferencia(s) con respecto a lo anterior. la(s) diferencia(s) del conjunto de datos anterior. Si se trata de un test, interprétalo de forma clásica e interpretarlo de forma clásica y volver a analizar la situación con una segunda mirada y luego luego tenga en cuenta las afirmaciones hechas aquí en el libro. ¿Hay diferencias y si es así, ¿cuáles?

Como ejemplo de la práctica de R, el conjunto de datos anterior (s. cap. 4.4.10) de Dobson (1990) sobre la medición del peso de las plantas se analiza con más detalle en un diseño de control frente a un diseño de tratamiento (`ptII_quan_classicstats_varianceh omogeneity.r`). Utilizamos el modelo lineal modelo lineal `lm.fit` y añadimos un `spreadLevelPlot` con `spreadLevelPlot()` de `car` (véase la Fig. 4.61).

```
> # variance homogeneity/ Heteroscedasticity
> # levene test
> car::leveneTest(weight, group)
Levene's Test for Homogeneity of Variance (center = median)
Df F value Pr(>F)
```

```
group 1 0.6203 0.4412
18
> # breusch-pagan test
> lmtest::bptest(lm.fit)
studentized Breusch-Pagan test
data: lm.fit
BP = 1.1565, df = 1, p-value = 0.2822
> # non-constant error variance
> spreadLevelPlot(lm.fit)
Suggested power transformation: 3.391163
> # non-constant variance score test
> # also breusch-pagan test, extensions by cook and weisberg
> ncvTest(lm.fit)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.8935197, Df = 1, p = 0.34453
> # identical to
> # studentize = logical. If set to TRUE Koenker's studentized
> # version of the test statistic will be used.
> lmtest::bptest(lm.fit, studentize=FALSE)
Breusch-Pagan test
data: lm.fit
BP = 0.89352, df = 1, p-value = 0.3445
```

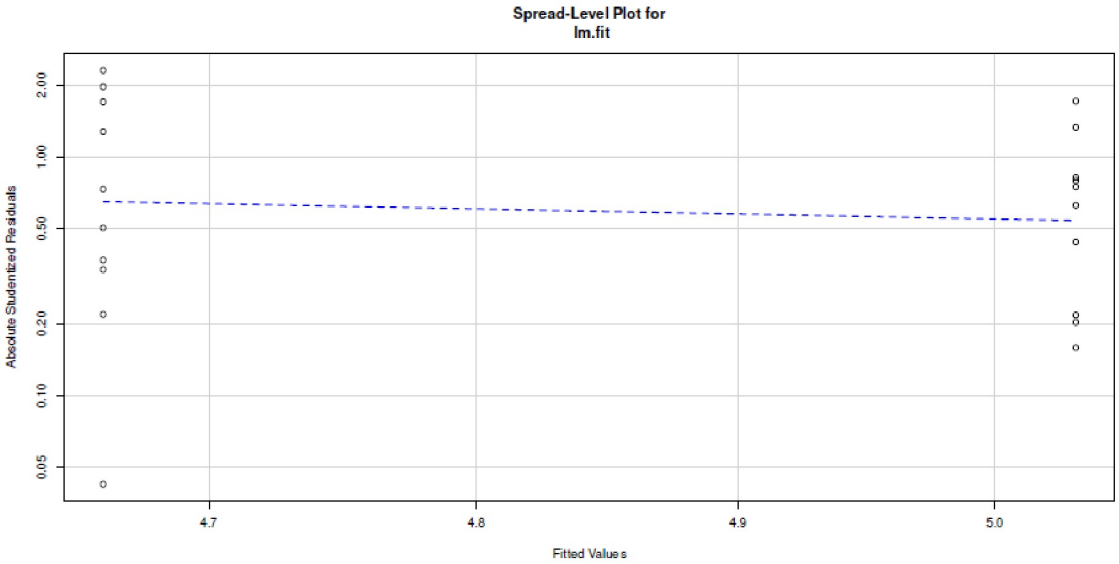


Figura 4.61. *Examen de la homogeneidad de las varianzas (`spreadLevelPlot()` de `car`)*

4.4.12 Valores atípicos y datos influyentes

La mayoría de los modelos son de naturaleza lineal – lo que sería motivo de una reflexión más profunda sobre si la linealidad per se es el modelo adecuado para el objeto de estudio de las ciencias sociales – y, por tanto, los valores atípicos pueden actuar como una palanca. Así, la pendiente de una línea de regresión puede cambiar considerablemente y, en consecuencia, distorsionar el poder explicativo y predictivo de las variables independientes. Pero los valores atípicos, es decir, los puntos extremos de los datos, bien pueden ajustarse al modelo o no. Esto sería objeto de una comprobación del modelo. Por ejemplo, puede haber valores atípicos naturales que no cuestionen en absoluto el modelo subyacente, sino que, de hecho, lo apoyen (véase la Fig. 4.62) y, en consecuencia, no deban eliminarse del conjunto de datos.
(`ptII_quan_classicstats_outliers--and-influentialpoints.r`).

```
# outlier and influential data points R-Code
seed <- 9876
set.seed(seed)
a <- c(1:20,100)
# add some noise
b <- a + rnorm(21,mean=5,sd=2)
data.frame(a, b=round(b,2))
```

Como puede verse en la Figura 4.62, el valor atípico podría encajar bien en el modelo general y, por tanto, no lo distorsiona. Con una línea de regresión recta y una línea de regresión local la imagen cambia sólo ligeramente.

```
# define outlier 21 R-Code
weg21 <- c(21)
# simple plots
plot.outlier(a=a, b=b, weg=weg21)
```

La estimación de ambos modelos lineales (con y sin valores atípicos, es el punto de datos 21) proporciona:

```
> # with outlier w21, without leverage point w22
> lmfit1 <- lm(b~a)
> summary(lmfit1)
Call:
lm(formula = b ~ a)
Residuals:
Min 1Q Median 3Q Max
-3.6639 -0.8790 -0.3158 1.1777 5.0269
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.87580 0.63802 7.642 3.29e-07 ***
a 1.01965 0.02577 39.564 < 2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 2.347 on 19 degrees of freedom
Multiple R-squared: 0.988, Adjusted R-squared: 0.9874
F-statistic: 1565 on 1 and 19 DF, p-value: < 2.2e-16
> # without any outlier w21 or leverage point w22
> lmfit2 <- lm(b[-c(weg21)]~a[-c(weg21)])
> summary(lmfit2)
Call:
lm(formula = b[-c(weg21)] ~ a[-c(weg21)])
Residuals:
Min 1Q Median 3Q Max
-3.4863 -1.4454 -0.3174 1.3495 4.2866
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.03057 1.09291 3.688 0.00168 **
a[-c(weg21)] 1.10310 0.09123 12.091 4.47e-10 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 2.353 on 18 degrees of freedom
Multiple R-squared: 0.8904, Adjusted R-squared: 0.8843
F-statistic: 146.2 on 1 and 18 DF, p-value: 4.469e-10
> coef(lmfit1)
(Intercept) a
4.875803 1.019653
> coef(lmfit2)
(Intercept) a[-c(weg21)]
4.030570 1.103104
```

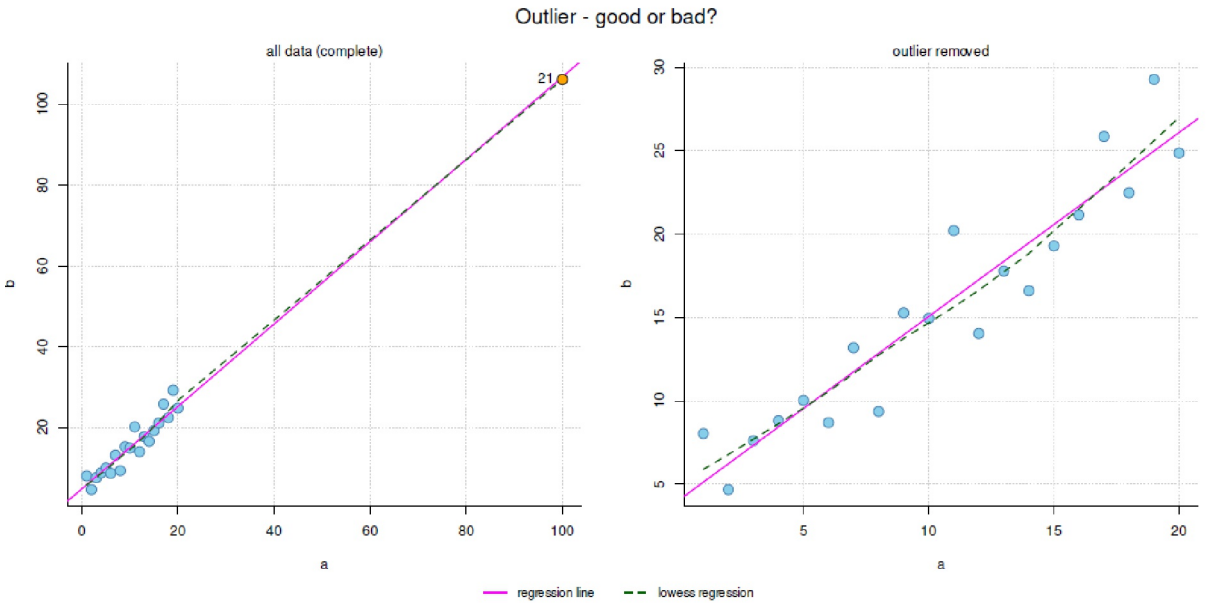


Figura 4.62. *Valores atípicos (con y sin punto de datos 21, línea de regresión lineal y local)*

Las estadísticas descriptivas proporcionan

```
# descriptive statistics with and without outlier 21
# with outlier
describes(data.frame(a,b))
summary(data.frame(a,b))
# without outlier 21
describes(data.frame(a[-weg21],b[-weg21]))
summary(data.frame(a[-weg21],b[-weg21]))
```

La prueba t muestra diferencias drásticas dependiendo de si se tiene en cuenta que el vector b se creó a partir del vector a y si la prueba se considera dependiente o no:

```
# t-test unpaired with outlier
t.test(a,b)
# t-test unpaired without outlier 21
t.test(a[-weg21],b[-weg21])
# t-test paired with outlier
t.test(a,b,paired=TRUE)
# t-test paired without outlier 21
t.test(a[-weg21],b[-weg21],paired=TRUE)
```

Por lo tanto, ignoramos los resultados de la prueba *t* no emparejada. Sin embargo, es importante ver cómo influyen en los resultados y en las conclusiones los supuestos previos, es decir, la independencia o dependencia de las muestras, y la eliminación de los valores atípicos. En el caso de la correlación muestra cómo el valor atípico estabiliza todo el modelo y reduce la incertidumbre:

```
# correlation with outlier
cor.test(a,b)
# correlation without outlier 21
cor.test(a[-weg21],b[-weg21])
```

El paquete R `car` proporciona una función R para evaluar los valores atípicos en los residuos `outlierTest()`, y una función R `leveragePlots()` para evaluar el apalancamiento de los puntos de datos

que corresponde aproximadamente al código R anterior (gráfico de dispersión y líneas de regresión, véase Fig. 4.63 arriba con todos los datos y abajo sin el punto de datos 21).

```
> # outlier test - for residuals
> outlierTest(lmfit1)
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
rstudent unadjusted p-value Bonferroni p
19 2.476112 0.023442 0.49228
> outlierTest(lmfit2)
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
rstudent unadjusted p-value Bonferroni p
19 2.184596 0.043213 0.86426
> # leverage plots
> leveragePlots(lmfit1, panel.first=grid())
> leveragePlots(lmfit2, panel.first=grid())
```

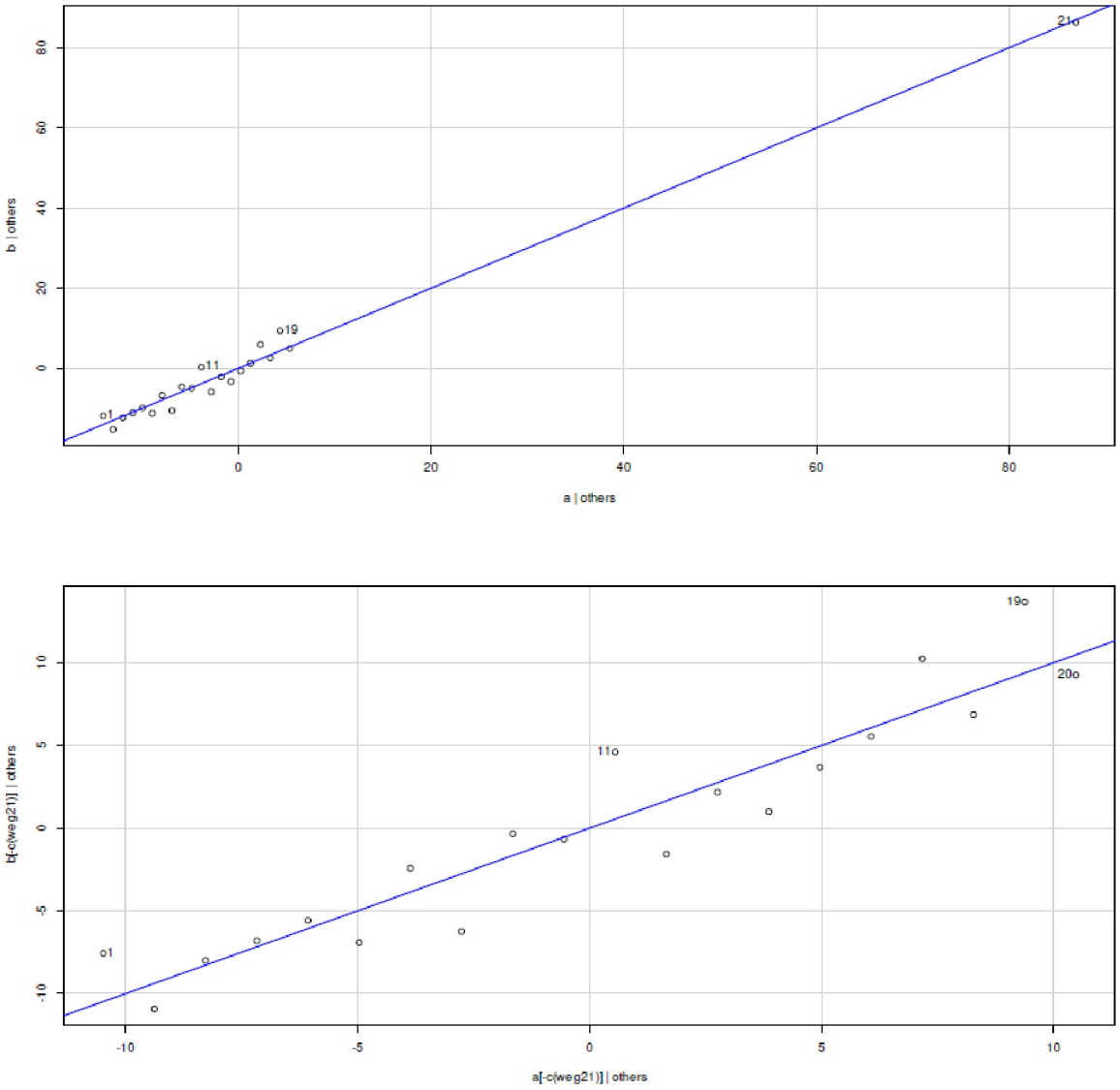


Figura 4.63. *Valores atípicos (con y sin punto de datos 21, gráfico de palanca)*

Hasta ahora, se ha demostrado que un valor atípico no significa necesariamente que el modelo esté sesgado. En el ejemplo anterior, el valor atípico incluso estabiliza el modelo. El siguiente código R muestra

cómo un punto de datos cuestiona el modelo de forma masiva, aunque a primera vista parezca menos dramático que el punto de datos anterior. Pero esto es engañoso, como la Figura 4.64 muestra.

```
# now with a real leverage point, "not an outlier"
a.LP <- c(a,80)
b.LP <- c(b,10)
ab.l <- length(a.LP)
# define leverage point 22
weg22 <- c(22)
# simple plots
plot.outlier(a=a.LP, b=b.LP, weg=weg22)
```

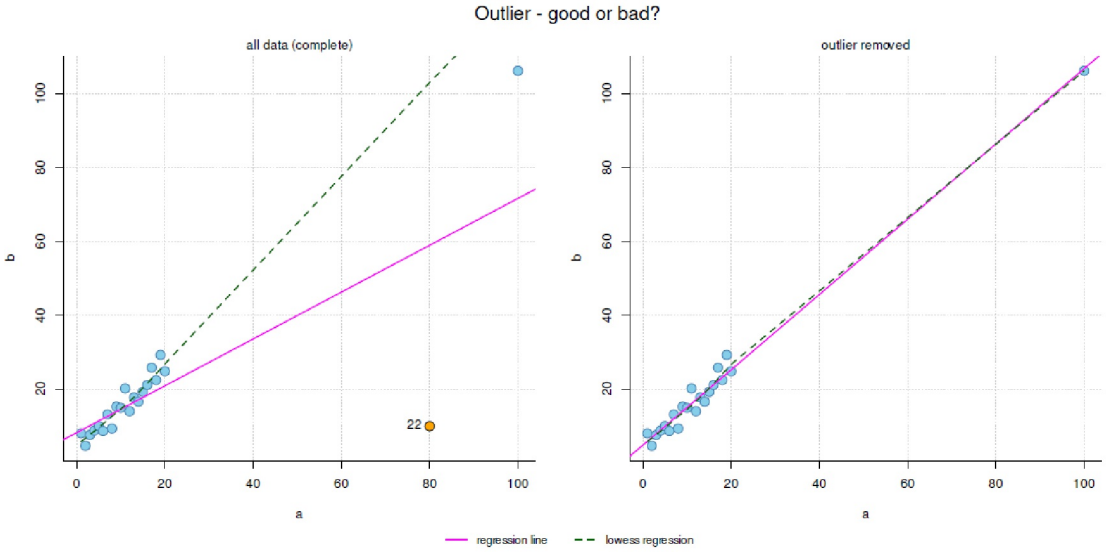


Figura 4.64. Datos influyentes (con y sin punto de datos 22, línea de regresión lineal y local)

Podemos trazar esto en paralelo para comprender mejor la influencia respectiva:

```
# plot various outliers and their influeces
# par(oma=c(2,1,1,1), "cex.axis"=1, bty="l", mfrow=c(1,1))
dev.off()
plot(a.LP, b.LP, col="steelblue", panel.first=grid(), pch=21,
          cex=1.4, bg="skyblue", bty="l", main="", xlab="a", ylab="b")
# outlier
points(a.LP[weg21],b.LP[weg21], col="black", pch=21,
          cex=1.4, bg="green")
# leverage point
points(a.LP[weg22],b.LP[weg22], col="black", pch=21,
cex=1.4, bg="red")
# mark with text
text(a.LP[c(weg21,weg22)],b.LP[c(weg21,weg22)],c("21","22"), pos=1)
# complete with w22 outlier and w22 leverage point
lmfit3 <- lm(b.LP~a.LP)
summary(lmfit3)
abline(lmfit3,col="purple",lty=1, lwd=1)
# with outlier w21, without leverage point w22
abline(lmfit1,col="green",lty=2, lwd=1)
# without any outlier w21 or leverage point w22
abline(lmfit2,col="skyblue",lty=2, lwd=1)
# without outlier w21, with leverage point w22
lmfit4 <- lm(b.LP[-c(weg21)]~a.LP[-c(weg21)])
summary(lmfit4)
abline(lmfit4,col="red",lty=2, lwd=1)
```

```
mtext("Outliers and different influences", outer=TRUE, line=-2,
          cex=1.5, side=3)
par(fig=c(0,1,0,1), oma=c(1,0,0,0), mar=c(0,0,0,0), new=TRUE)
plot(1, type="n", bty="n", xaxt="n", yaxt="n")
legend("bottom", legend=c("complete", "w/o [22]",
          "w/o [21]", "w/o [21,22]"),
lty=c(1,2,2,2), lwd=1, col=c("purple","green","red","skyblue"),
          bty="n", cex=.9, xpd=TRUE, horiz=TRUE)
```

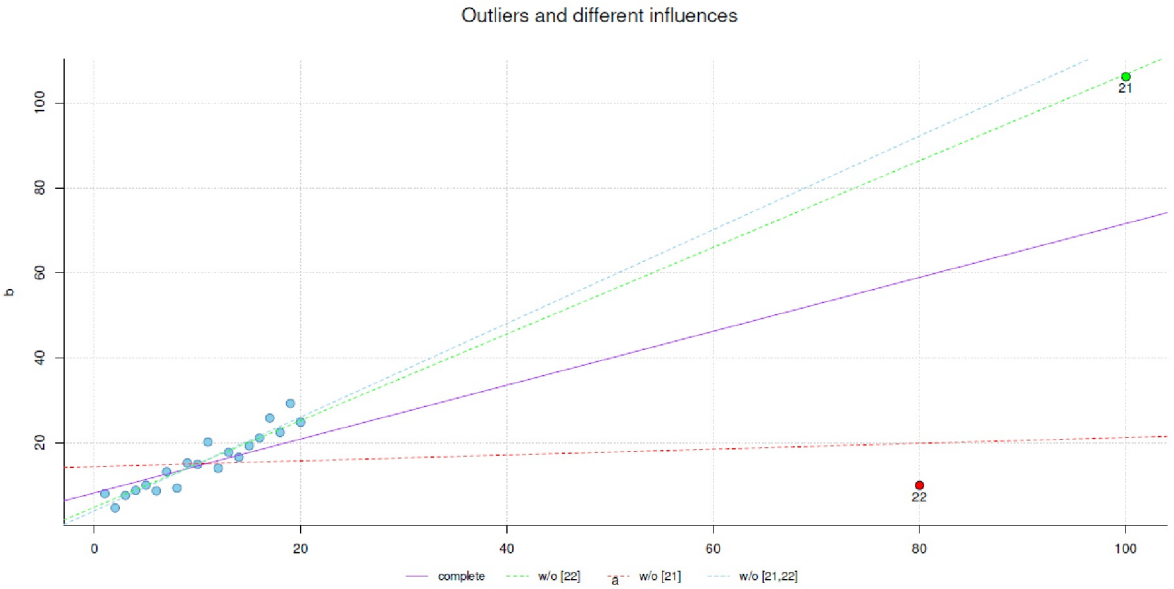


Figura 4.65. *Valores atípicos y puntos de datos influyentes (con línea de regresión, los cuatro casos)*

En la Figura 4.65 se pueden ver cuatro líneas de regresión, que muestran fácilmente cómo un solo punto de datos puede desarrollar un efecto palanca. La línea sólida roja corresponde a la regresión con el punto de datos (= DP) DP21 (véase más arriba, marcado en verde), pero sin el nuevo punto de apalancamiento (DP22, marcado en rojo). Las líneas discontinuas marcan los cambios en la regresión: [1] sin DP21 (magenta), [2] con DP22 (verde mar), así como [3] sin DP21 pero con DP22 (azul cielo). El caso [1] muestra un ligero cambio, aunque la estabilización por DP21 ya es visible. El caso [2], en cambio, muestra un gran cambio en la pendiente de regresión debido a la adición de DP22. Si ahora se suprime DP21, surge el caso [3] y la línea de regresión cambia completamente de su cambia de dirección y viene determinada principalmente por DP22. Un vistazo a los modelos lineales y sus coeficientes lo confirma.

```
# compare models with each other
# beta coefficient, but also intercept
comp.beta <- rbind(lmfit3$coef, lmfit1$coef, lmfit4$coef, lmfit2$coef)
rownames(comp.beta) <- c("complete", "w/o [22]",
          "w/o [21]", "w/o [21,22]")
colnames(comp.beta) <- c("(Intercept)","beta")
comp.se <- rbind(display(lmfit3)$se, display(lmfit1)$se,
          display(lmfit4)$se, display(lmfit2)$se)
rownames(comp.se) <- c("complete", "w/o [22]",
          "w/o [21]", "w/o [21,22]")
colnames(comp.se) <- c("(Intercept)SE","beta(SE)")
```

Aquí como salida las comparaciones para los coeficientes del modelo y los errores estándar:

```
> comp.beta
(Intercept) beta
complete 8.230809 0.63437538
w/o [22] 4.875803 1.01965270
w/o [21] 14.396353 0.06875783
```

```
w/o [21,22] 4.030570 1.10310379
> comp.se
(Intercept)SE beta(SE)
complete 3.6936183 0.12480228
w/o [22] 0.6380217 0.02577246
w/o [21] 2.0081903 0.09558168
w/o [21,22] 1.0929083 0.09123422
```

Y la función `outlierTest()` encuentra fácilmente los puntos de datos relevantes – hablando estadísticamente:

```
> # complete
> # complete with w21 outlier and w22 leverage point
> outlierTest(lmfit3)
    rstudent unadjusted p-value   Bonferroni p
22 -26.071341 2.4523e-16          5.3950e-15
21   7.811528 2.3837e-07          5.2441e-06
>
> # w/o [21]
> # without outlier w21, with leverage point w22
> outlierTest(lmfit4)
    rstudent unadjusted p-value   Bonferroni p
21 -12.12905 4.2473e-10           8.9194e-09
```

Observe que en la salida de `lmfit4`, el valor atípico `DP22` se denomina `DP21`. La razón es que el `DP21` original se excluyó antes del análisis, de modo que `DP22` se convirtió en `DP21`. Con `lmfit3`, por otra parte, es importante observar que la prueba identifica no sólo `DP22` sino también `DP21` como valores atípicos estadísticamente significativos por convención. Dado que aquí los datos fueron autogenerados y, por tanto, está teóricamente claro que `DP21` es un componente genuino del modelo, esto demuestra a la inversa que no hay que creerse sin más este tipo de pruebas, sino comprobar cuidadosamente la validez de contenido de cada resultado. Los gráficos de la Figura 4.65 muestran que `DP21` tiene un efecto estabilizador del modelo y es un valor atípico, pero no uno que distorsione el modelo subyacente. `DP22`, en cambio, distorsiona el modelo. En un entorno real, `DP21` se dejaría en el modelo, mientras que `DP22` podría eliminarse tras un análisis detallado de las razones por las que se produjo esta fecha. Esto iría precedido de una aclaración de la legitimidad de esta fecha, dos veces, para no perder prematuramente información de forma inadecuada.

El análisis de los modelos lineales (véase más arriba) muestra que, en función del caso analizado, los coeficientes de la intercepción del eje Y y de la pendiente difieren considerablemente. Lo mismo ocurre con los errores estándar respectivos. Por ejemplo, el modelo original (caso [1]) tiene una pendiente 14 veces mayor que el caso [4] (1.02/0.07 = 14.57) y en el caso [4] el error típico es mayor que el coeficiente estimado (valor t= 0.07/0.1 = 0.7). Así pues, las conclusiones de contenido conducen a resultados diametralmente opuestos. Aquí merece la pena observar repetidamente la figura gráfica 4.65, que hace que este comportamiento sea intuitivamente accesible.

Una reacción común ante los valores atípicos es simplemente excluir los puntos de datos extremos o muy inusuales del conjunto general. Antes de hacerlo, es necesario realizar un extenso análisis combinado gráfico-numérico y de contenido, como se acaba de explicar. La pregunta es: ¿se trata de un dato erróneo, de un error de medición, de un acontecimiento raro pero posible y qué significan en teoría tales valores atípicos? Si la explicación teórica no se intenta de acuerdo con una investigación causal determinada, las conclusiones pueden ser inadecuadas. En cualquier caso, habría que preguntarse si, siguiendo la línea argumental repetidamente citada de Gelman y sus colegas, un modelo ampliado y más complejo sería una mejor opción, o al menos un segundo modelo, si el alcance del modelo principal termina en un determinado intervalo de datos. Por ejemplo, puede ser que un instrumento de encuesta ya no diferencie suficientemente. Por ejemplo, los tests de cociente intelectual convencionales ya no diferencian suficientemente entre la área más baja y más alta. Per eso se necesitan instrumentos especiales para examinar personas en la zona de inferioridad o superdotación.

A la inversa, sin embargo, es posible que los supuestos valores atípicos sean puntos de datos influyentes, pero que se ajustan plenamente al modelo y sólo aparecen como valores atípicos porque la muestra concreta entre ellos y la masa de puntos de datos no contiene otros puntos de datos por casualidad (véase la Fig. 4.62).

Por lo general, los requisitos previos de las pruebas se comprueban con pruebas estadísticas antes de aplicar el procedimiento de prueba elegido. Estas pruebas preliminares se basan de nuevo en la estadística clásica y en un valor p, es decir, en la lógica de significación clásica (por ejemplo, pruebas de distribución normal,homogeneidad de la varianza, etc.). Desde hace algún tiempo, estas pruebas de modelos se complementan cada vez más o incluso se sustituyen por pruebas de modelos gráficos. Aquí el usuario se encuentra de nuevo con el conocido problema de dónde fijar el punto de corte a partir del cual una violación sustancial de los prerequisitos de la prueba se considera plausible y dónde sólo *es aún posible*. Todo ello tiene consecuencias, a saber, la legitimación del procedimiento de análisis de datos dirigido o el paso a otro nivel de información, es decir, la reducción de escala. Además, los procedimientos robustos pueden (Wilcox, 2012; CRAN, 2018 para implementaciones en R) realizar transformaciones de datos, eliminar puntos de datos, etc. En todas estas posibilidades subyace el hecho de que tienen consecuencias e influyen en los resultados – recordemos: no hay análisis de datos sin modificación directa de los mismos. En cada caso concreto, es necesario examinar qué consecuencias de carácter sustantivo surgen y hasta qué punto parecen plausibles calculando y comparando modelos con y sin estos cambios introducidos. Si, por ejemplo, está claro que un dato extremo se ha producido incorrectamente debido a una medición o codificación incorrecta, entonces la exclusión es legítima. Si, por el contrario, no hay indicios de ello, sería negligente eliminar sin más un dato sin una explicación plausible al nivel de su contenido.

Así pues, el uso de procedimientos gráficos, que también se utilizan en el AED (Tukey, 1977), parece ser mejor que las pruebas de significación. Por ejemplo, representar gráficamente los datos empíricos frente a los de la distribución normal teórica o, en el caso de los modelos lineales, hacer lo mismo para la evaluación de los residuos es un método de probada eficacia. Lo mismo se practica en la estadística bayesiana para la evaluación de las distribuciones de probabilidad posteriores para investigar si están distribuidas de forma sesgada o simétrica. Los métodos gráficos, a diferencia de los coeficientes, permiten un enfoque intuitivo que tiene en cuenta no sólo los estadísticos de resumen, sino la representación de todos los puntos de datos. Se ofrecen ejemplos en R en Vanderplas y Hofmann (2015-08), Loy, Follett y Hofmann (2015), Loy, Hofmann y Cook (2016) y Loy y Hofmann (2013).

En sentido estricto, si no se cumplen los requisitos de la prueba, esta no puede utilizarse. Las excepciones son los valores empíricos sobre la solidez con la que reaccionan los procedimientos a las violaciones de sus requisitos previos, pero estos solo pueden objetivarse hasta cierto punto. Entonces, ¿cuáles son las consecuencias en un caso concreto si se utiliza un procedimiento a pesar de las violaciones de las condiciones previas? Esto sitúa la justificación en las proximidades de las decisiones subjetivas, ya que en los casos límite faltan criterios vinculantes. La descripción de resultados gráficos, la discusión de puntos concretos en el contexto del diseño de la investigación, etc., están muy próximos a los métodos reconstructivos de la investigación cualitativa. En este sentido, existe una similitud estructural entre la argumentación cualitativa y los múltiples procesos de toma de decisiones basados en el análisis estadístico. Así, inicialmente se puede hacer poco con la observación de que un punto de datos se aleja demasiado de lo esperado. De ello se derivan preguntas como "¿Qué significa esto?", "¿Qué consecuencias resultan si se omite el dato o no? – y ya comienza una discusión cualitativa sobre cantidades realmente numéricas (Gürtler y Huber, 2006). ¿Puede llamarse objetivos los estudios puramente cuantitativos? Eso no nos parece justificado, dada la abundancia de decisiones no objetivables en el curso de estos estudios.

Un método habitual en ausencia de requisitos de prueba es reducir el nivel de escala y continuar con un procedimiento equivalente al nivel de escala reducido. Pero incluso entonces hay requisitos previos que comprobar, aparte del hecho de que una reducción del nivel de escala conlleva una pérdida de información. Cambiar el nivel de escala también requiere un cambio en la formulación de la hipótesis o la pregunta. Por ejemplo, la pregunta ya no se refiere a las diferencias medias en una escala original (por ejemplo, cuestionario), sino ahora a las diferencias en el nivel de clasificación (rangos). La diferencia cualitativa en las clasificaciones, que está contenida en el nivel de intervalo, se omite. Sin embargo, es posible utilizar tanto el método original como el reducido y comparar los resultados entre sí. Si son iguales o conducen a conclusiones idénticas, el procedimiento original puede utilizarse con precaución y es necesaria una explicación en la presentación posterior de los resultados. En el informe debe retomarse el problema y discutirse sus dificultades y consecuencias.

Además de estas soluciones radicales y a menudo inapropiadas, se necesitan más libros de texto que muestren cómo los investigadores pueden crear sus propias soluciones de forma creativa pero científicamente

seria. A veces no se necesita una gran estadística (inferencial), como muestran los estudios de casos sobre el potencial de reintegración en el contexto de la terapia de la adicción (véase el cap. 5.5.7) o sobre la supervivencia en el Titanic (véase el cap. 5.5.7). Allí, los resultados de las consideraciones cualitativas se utilizan en combinación con el análisis cuantitativo-exploratorio de datos, sin por ello dotarlos de menos potencial para una posible pretensión de validez.

Tras las excursiones por el mundo de la significación, los valores p y el sesgo, siguen ahora las reflexiones sobre cómo tratar el tamaño y la frecuencia de los efectos y que la referencia a la escala original no debe perderse en el proceso.

4.4.13 Tamaño del efecto, frecuencia y relación con la escala original

La investigación vive con múltiples compromisos o soluciones intermedias. Muchas de las variables influyentes que intervienen en un análisis de potencia a priori no se pueden seleccionar libremente en los estudios empíricos. Las muestras suelen caracterizarse por limitaciones (por ejemplo, falta de tiempo, disponibilidad de personas y recursos financieros). Para la selección de la probabilidad crítica de superación (error de tipo I) rara vez hay razones razonables y a menudo la decisión se toma en función de la corriente dominante (por ejemplo, 5%). También en el caso del error de tipo II es difícil hacer una declaración clara de antemano sobre el tamaño objetivo, excepto, por ejemplo, seguir el consejo de Cohen (1969) de elegir el error de tipo II unas cuatro veces mayor que el error de tipo I. Esto se parece más bien a una regla empírica más que una solución apropiada para una cuestión empírica concreta. Pero esto no debe achacarse a Cohen, que se refiere a metaestudios y a una medida de comparabilidad entre estudios y no hace afirmaciones sobre todas las escalas originales posibles en un contexto concreto. En contextos diferentes, estos límites tienen que redefinirse y justificarse de todos modos, porque la investigación no se desarrolla (únicamente) en el metanivel descontextualizado. Hattie (2009) ha presentado un metaestudio muy amplio "Visible Learning" que muestra que los límites trazados metaanalíticamente son más suaves en la educación. Szucs y Ioannidis (2017) utilizan un metaestudio de artículos de neurociencia cognitiva y psicología para mostrar cómo las muestras pequeñas que se utilizan habitualmente allí afectan a la potencia del efecto. Los mismos tamaños de efecto "libres de contexto" tienen efectos diferentes según el contexto.

Tabla 4.13. *d de Cohen (Cohen, 1969)*

Interpretación	r de Cohen (Correlación Pearson)	d de Cohen
ningún efecto	$0 < r < 0.1$	$0 < d < 0.2$
bajo	$0.1 < r < 0.3$	$0.2 < d < 0.5$
medio	$0.3 < r < 0.5$	$0.5 < d < 0.7$
alto	$0.5 < r$	$0.7 < d$

Cohen (1992) ofrece recomendaciones aproximadas para estimar la *d* de Cohen en las ciencias sociales (véase la tabla 4.13) con el fin de generar una variable independiente de la escala para las comparaciones entre estudios. Inicialmente, se hace una distinción gratificantemente sencilla entre "ningún efecto", "bajo", "medio" y "alto". Cuatro categorías parecen ser un compromiso factible – entre una gran precisión y una adaptabilidad razonable. De este modo, resulta más fácil determinar las tendencias centrales que con gradaciones demasiado finas que, además, dependen de la muestra y se sobrecargan rápidamente como consecuencia de una interpretación excesiva. Sin embargo, son más importantes la escala original de los efectos relevantes (dirección y nivel) y su justificación sustantiva. Para interpretaciones extensas, una medida abstracta como la *d* de Cohen no es suficiente, ya que sólo tiene una importancia secundaria, especialmente cuando se trata de comparar varios estudios, pero no en relación con la escala original. Si se dispone de una escala original, ésta siempre desempeña el papel relevante – en comparación con una variable descriptiva

independiente de la escala. Para otras medidas del tamaño del efecto, como el coeficiente de correlación r de Pearson, el coeficiente de determinación R^2, así como *g* de Hedge, Δ de Glass, f^2 de Cohen, η^2 (parcial), ω^2, Ψ , etc., se aplica en principio la misma línea de razonamiento. Incluso los cálculos de potencia a priori deben denotar primero con mucha precisión un efecto sobre la escala antes de traducirlo en una medida independiente de la escala. Los tamaños del efecto no sólo están disponibles para diferencias de medias estandarizadas, sino también para modelos lineales (ANOVA, η^2, ω^2), categóricos (riesgo relativo = RR, diferencia de riesgo = RD, ω y *h* de Cohen u odds ratio), datos ordinales (*d* de Cliff) e incluso para estudios caracterizados por un alto rendimiento en la selección de datos (factor Z), por ejemplo en la investigación bioquímica y farmacéutica. En principio, pueden derivarse intervalos de confianza clásicos para estas cantidades.

> Tarea 4.16: Aumentar el tamaño de la muestra
>
> El script de R `pv.vs.cd()` permite además elegir siempre el mismo valor inicial para las muestras crecientes. Todo lo que se necesita es la llamada `pv.vs.cd(usesameseed=TRUE)`. ¿Qué cambia ahora en comparación con la llamada que genera valores de salida diferentes en cada caso?

El tamaño del efecto per se tiene poco que ver con los *p*-valores, ya que se puede demostrar fácilmente que un estudio con un tamaño de muestra grande y tamaños de efecto pequeños puede llegar a los mismos *p*-valores que un estudio con un tamaño de muestra pequeño y un tamaño de efecto grande (Hubbard & Lindsay, 2008). No es difícil demostrar mediante R-script que la disminución constante de los valores *p* se debe únicamente al aumento del tamaño de la muestra, mientras que, al mismo tiempo, la potencia del efecto permanece constante en el contexto de las fluctuaciones aleatorias (véase la Fig. 4.66 con diferentes valores de salida para el generador aleatorio, abajo con el mismo valor inicial, `ptII_quan_classic-stats_effectsizes.r`):

```
# call function to relate N, pv and Cohen's d
# without the same seed
res.notsameseed <- pv.vs.cd()
# with the same seed
res.sameseed <- pv.vs.cd(usesameseed=TRUE)
```

Como puede observarse, el coeficiente de pendiente para el modelo con los valores *p* tiene un claro sesgo negativo, es decir, un tamaño de muestra creciente se asocia sistemáticamente con valores *p* más pequeños. Por el contrario, el coeficiente de pendiente para el modelo *d* de Cohen es prácticamente cero y el modelo se describe casi exclusivamente por el intercepto. Mientras que con muestras muy pequeñas el tamaño del efecto fluctúa fuertemente, se nivela con el aumento del tamaño de la muestra y varía alrededor de $d = 0.1$. Ignoramos aquí si los modelos lineales son realmente los mejores modelos. Especialmente para el modelo con los tamaños del efecto, un modelo no lineal sería probablemente la mejor opción debido a las grandes fluctuaciones iniciales. Sin embargo, dado que especificamos y conocemos los valores teóricos, resulta que una pendiente casi horizontal describe el proceso subyacente a los datos de forma bastante adecuada, teniendo en cuenta las fluctuaciones aleatorias. Por tanto, el coeficiente de determinación extremadamente bajo no es relevante. Una mirada al gráfico y el conocimiento de las relaciones estructurales entre los datos son más prometedores que la fe ciega en los coeficientes.

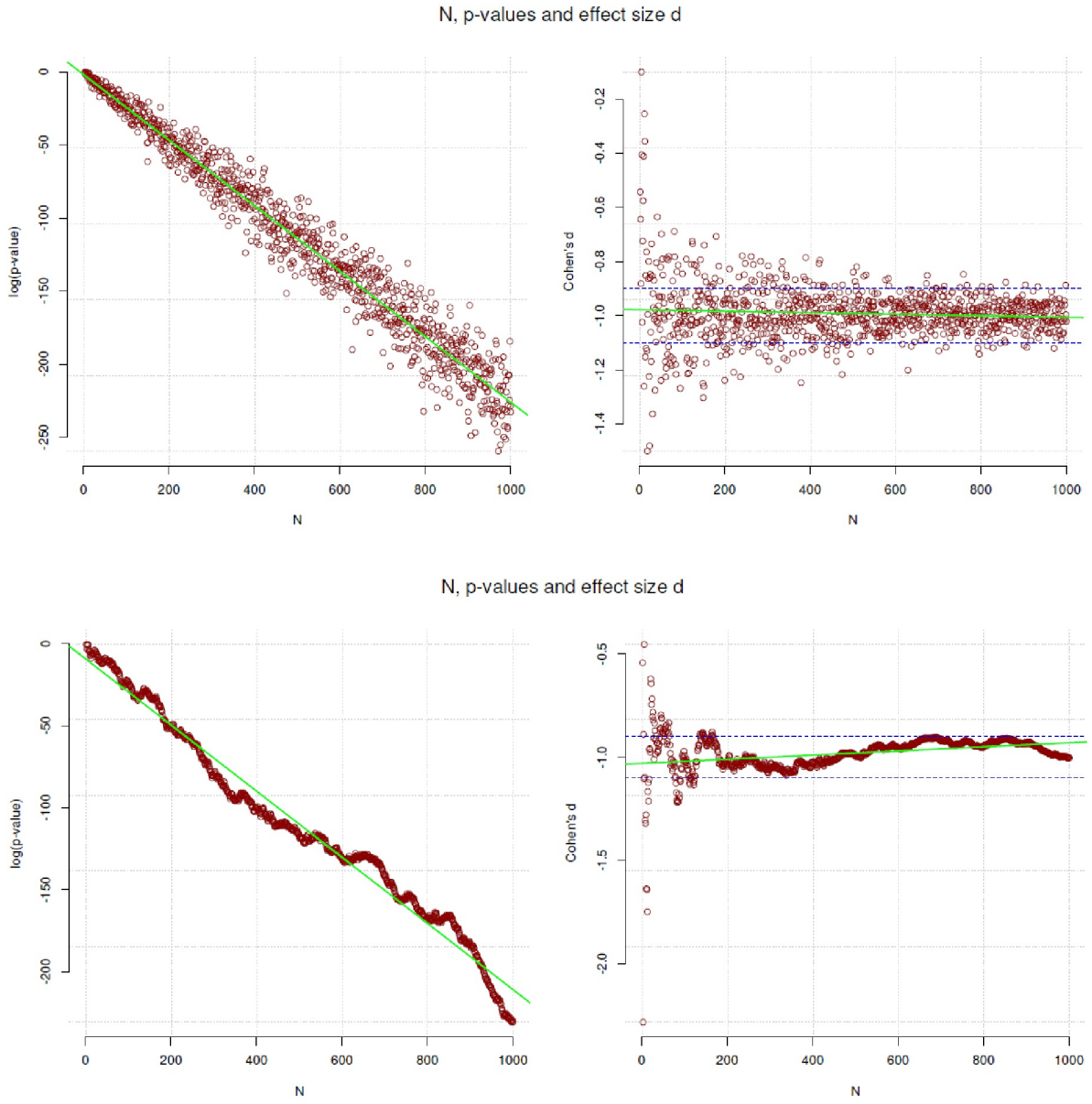


Figura 4.66. *Relación entre tamaño de muestra, valor p y d de Cohen*

Debido a la dependencia directa de la muestra, los valores *p* no pueden considerarse una medida objetiva de las pruebas a favor o en contra de una hipótesis. Hubbard y Lindsay (ibíd., p.77) señalan al respecto: "Contrariamente a lo que afirma Fisher, el valor p no es una medida objetiva de las pruebas en contra de una hipótesis [...]". Lamentablemente, puede observarse que, al igual que ocurre con los factores de Bayes (véase el capítulo 6.8.1.4), hay bastantes autores que complican y solidifican la jerarquía de Cohen, que sigue siendo bastante sencilla y tiene sentido por estas razones, y propagan interpretaciones completamente libres de contexto de determinados rangos de valores para los distintos tamaños del efecto. Una visión general es proporcionada, por ejemplo, por la viñeta para el paquete R `effectsize`, apropiadamente llamado "Interpretación automatizada de índices de tamaño del efecto" – recordamos a más tardar ahora la advertencia de Gigerenzer (2018) y Gigerenzer, Krauss & Vitouch (2004) sobre el tema de las *interpretaciones automatizadas*. Además de la idealización de las inferencias automatizadas, parece haber una necesidad de libros de cocina en el campo metodológico en lugar de construir conjuntos que deben aplicarse de manera flexible y sensible con sentido común al contexto.

Fisher tampoco comentó otra cosa hacia el final de su vida (véase el capítulo 4.3.2) con su obvio golpe de costado en dirección a Neyman-Pearson.

Aunque es cierto que muchas de estas directrices se concibieron originalmente como reglas empíricas, por desgracia la práctica es que a menudo se interpretan estrictamente según estas reglas empíricas, comparable al manejo de los valores *p*. Sin embargo, una regla empírica denota una estimación y un valor empírico, que naturalmente cambia con nuevas experiencias y estimaciones y no reside independientemente de la situación. La confusión de las reglas empíricas con las leyes hace que la discusión se desplace al nivel de estos tamaños del efecto abstractos y libres de contexto, privados de su significado original, sin que los investigadores emprendan el camino, ciertamente arduo pero provechoso, de vuelta a la escala original. Sin embargo, esto es necesario para poder ofrecer una interpretación razonable en la línea de la pregunta de investigación planteada. Por supuesto que merece la pena fijarse en variables abstractas y libres de contexto cuando se van a comparar estudios que no tienen exactamente el mismo diseño y, por tanto, a menudo no es posible realizar una comparación directa en las escalas originales, es decir, para casi todas las comparaciones y especialmente para los metaestudios. Sin embargo, esto no exime de la obligación de reflexionar detenidamente sobre el significado real de una cantidad hallada empíricamente en el contexto estudiado y las conclusiones que de ello se derivan. Estandarizar significa simplemente situar las cantidades en una escala técnicamente comparable, pero no situarlas en una escala comparable en términos de contenido y, desde luego, no interpretarlas de una manera libre de contexto y de leyes.

Por ejemplo, podríamos plantearnos estudiar *estados* como las fluctuaciones emocionales o motivacionales a corto plazo. Por otro lado, investigamos simultáneamente cambios y *rasgos de personalidad* a largo plazo. Si eliminamos las escalas y comparamos los resultados sólo a nivel abstracto, ¿podemos afirmar que un "efecto medio" aquí (para los estados) se corresponde con un "efecto medio" allí (para los rasgos) e incluso ¿podemos hacer este tipo de comparaciones? ¿Qué significaría tal comparación, concretamente o también en un metaestudio? Los estados, como su nombre indica, pueden cambiar a corto plazo y de forma espontánea, por ejemplo por una información, ver una película, escuchar música, etc. Con los rasgos de personalidad los cambios no parecen tan espontáneos. En películas la música y la información no bastan. Una empresa así requiere voluntad, perseverancia, tiempo, etc., es decir, mucho trabajo. Desde el punto de vista estadístico, una comparación no supone ningún problema y no se trata de eso. Sin embargo, si luego se consultan las escalas originales, un punto aquí significa algo muy distinto de lo que significa allí, y esto queda claro muy rápidamente debido a la orientación de una escala en función del contenido. Así pues, se trata de las conclusiones en el nivel del contenido y ahí, después de todos sus cálculos, la estadística no ha perdido nada. No es más que un dato entre muchos otros. No en vano Gigerenzer (2018) se pregunta: "¿Qué hacían los psicólogos antes de la revolución de la inferencia?". Al mismo tiempo, cita al matemático, físico e investigador inglés Sir Isaac Newton (1642-1726), que consideraba la estadística relevante para el control de calidad, pero no para la ciencia. En el capítulo 4.3.3 argumentamos de forma similar utilizando el ejemplo de la teoría de Neyman-Pearson, que es bastante excelente para el control de calidad. Sin embargo, dudamos si puede y debe jugar el mismo papel para la ciencia.

En R, existen varios paquetes R para los tamaños del efecto, como `psych`, `effsize`, `effectsize`, o `sjstats`, que también permiten transferir valores de una versión a otra. Mientras esto sea claramente posible en ambas direcciones, en realidad no podemos hablar de diferentes tamaños del efecto, sino sólo de diferentes perspectivas sobre el mismo fenómeno.

Un ejemplo construido del uso de potencias del efecto en R se parece a esto. En primer lugar se generan dos muestras y se comparan mediante una prueba *t*, se almacenan los valores *p* y se calcula la *d* de Cohen. (`ptII_quan_classicstats_effectsizes.r`).

```
> # different effect sizes, different n, approx. same p-values
> seed <- 9876
> set.seed(seed)
>
> n1 <- 80
> sigma1 <- 1.5
> s1 <- rnorm(n=n1, mean=10, sd=sigma1)
> s2 <- rnorm(n=n1, mean=10.4956458, sd=sigma1)
> t1 <- t.test(s1,s2)
> ES.v <- (mean(s2)-mean(s1))/sigma1
```

```
> ES.v
[1] 0.2057441
> cohensd(s1, s2, sd.theory=sigma1)
d|mean sd d|pooled sd d|theory sd
0.1897364 0.1897364   0.2057441
>
> n2 <- 1000
> sigma2 <- 10
> w1 <- rnorm(n=n2, mean=100, sd=sigma2)
> w2 <- rnorm(n=n2, mean=100.57567, sd=sigma2)
> t2 <- t.test(w1,w2)
> ES.w <- (mean(w2)-mean(w1))/sigma2
> ES.w
[1] 0.09921358
> cohensd(w1, w2, sd.theory=sigma2)
d|mean sd  d|pooled sd d|theory sd
0.10151269 0.10151269  0.09921358
```

El resultado de las pruebas y los cálculos muestra tamaños del efecto diferentes, pero valores *p* prácticamente idénticos para tamaños de muestra muy distintos. El tamaño pequeño de la muestra se compensa en términos de *p*-valores con un gran efecto, mientras que la muestra grande tiene un efecto pequeño, pero debido a la gran base de datos llega prácticamente al mismo *p*-valor. La cuestión de la significación es irrelevante en este caso. Se trata simplemente de demostrar las dependencias mutuas del valor *p*, el tamaño del efecto y el tamaño de la muestra.

```
> # compare values
> t1
Welch Two Sample t-test
data: s1 and s2
t = -1.2, df = 156.51, p-value = 0.232
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.8166086 0.1993763
sample estimates:
mean of x mean of y
10.12963 10.43825
> t2
Welch Two Sample t-test
data: w1 and w2
t = -2.2699, df = 1997, p-value = 0.02332
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1.8493259 -0.1349456
sample estimates:
mean of x mean of y
99.41472 100.40685
> all.equal(t1$pv, t2$pv)
[1] TRUE
> ES.v
[1] 0.2057441
> ES.w
[1] 0.09921358
```

Tarea 4.17: Tamaños de los efectos

Usando los códigos R anteriores, modifica la base de datos para obtener cualquier valor p que elijas (por ejemplo, p 0:42) para ambas pruebas t. Escriba un script R que se aproxime a esto. Puede trazarlo todo para simplificar. Compare los tamaños del efecto y los tamaños de muestra necesarios. Varía estos tamaños para tener una idea de sus dependencias. de sus dependencias.

En el paquete R `effsize` se pueden encontrar más valores del efecto

```
cliff.delta(w1,w2)
VD.A(w1,w2)
effsize:::cohen.d(w1,w2)
effsize:::cohen.d(w1,w2,hedges.correction=TRUE)
```

El paquete de R `effectsize` tiene potencias de efecto para modelos lineales

```
# lm and effect size
lm.w2.w1 <- lm(w2~w1)
display(lm.w2.w1)
eta_squared(lm.w2.w1)
eta_squared(lm.w2.w1, partial=FALSE)
eta_squared(lm.w2.w1, partial=FALSE)
epsilon_squared(lm.w2.w1)
omega_squared(lm.w2.w1)
effectsize:::cohens_f(lm.w2.w1)
```

y para convertir de una forma a otra

```
# transform from one size to the other
# just an example
lm.w2.w1.anova <- anova(lm.w2.w1)
# F value to eta^2
F_to_eta2(lm.w2.w1.anova$‘F value‘[1],
lm.w2.w1.anova$Df[1], lm.w2.w1.anova$Df[1])
# Cohens’ d to r
d_to_r(cohensd(w1,w2)[2])
# d_to_r <- function (d, ...) d/(sqrt(d^2 + 4))
```

Del mismo modo, las frecuencias y proporciones pueden examinarse en relación con el tamaño de sus efectos.

```
# effect size for proportions
ES.h(0.2,0.3)
# effect size for chi^2 GOF
prob.0 <-rep(1/3,4)
prob.1 <- c(0.4,rep((1-0.4)/3,3))
prob.0
prob.1
ES.w1(prob.0,prob.1)
```

Volvemos al ejemplo de Fisher de beber té del capítulo 4.3.2.1:

```
> # effect size for chi^2 association
> fisher.tea <- matrix(c(4,0,0,4), nrow=2,
                       dimnames=list(guess=c("milk","tea"),
                       truth=c("milk","tea")))
> fisher.tea
truth
guess milk tea
milk  4    0
tea   0    4
> ES.w2(prop.table(fisher.tea))
[1] 1
> fisher.wrong.tea <- matrix(c(3,1,1,3), nrow=2,
                       dimnames=list(guess=c("milk","tea"),
                       truth=c("milk","tea")))
> fisher.wrong.tea
truth
guess milk tea
milk  3    1
```

```
tea   1    3
> ES.w2(prop.table(fisher.wrong.tea))
[1] 0.5
```

y finalmente veremos `odds_to_rr()`, un ejemplo de HLM para convertir múltiples odds ratios en valores de riesgo relativo. Desafortunadamente, esta función `odds_to_rr()` ya no existe en las versiones más recientes de `sjstats`. Por lo tanto, la cargamos como una función auxiliar `odds_to_rr.v2()` ligeramente modificada, tomada de una versión anterior de `sjstats`. Los detalles pueden encontrarse en `ptII_quan_classicstats_effectsizes_helpfuncs_sjstats.r`. El conjunto de datos *Wells* del paquete R `carData` contiene la información, si los hogares de Bangladesh cambiaron el pozo que utilizaban. El conjunto de datos contiene cinco variables, véase `?Wells - switch` (change de un pozo inseguro a otro más seguro, sí/no), arsénico (contaminación por arsénico del pozo original, todo por encima de 0,5 se considera inseguro, en microgramos por litro), distancia (distancia al pozo seguro más cercano, en metros), educación (en años, cabeza de familia) y asociación (participación de los miembros de la familia en una organización local, sí/no).

```
# Data on whether or not households in Bangladesh
# changed the wells that they were using.
# http://www.stat.columbia.edu/~gelman/arm/examples/arsenic/wells.dat
?Wells
data(Wells)
summary(Wells)
wells.glm <- glm(switch~arsenic+distance+education+association,
                 data=Wells, family=binomial(link="logit"))
display(wells.glm)
# RR from glm
wells.glm.OR.RR <- odds_to_rr.v2(wells.glm)
```

A continuación se muestra la salida de `well.glm.OR.RR` que produce el riesgo relativo debido al `glm`

```
wells.glm.OR.RR
$OR
               OR       lower.ci   upper.ci
(Intercept)    0.8549505 0.7031210 1.0390475
arsenic        1.5952358 1.4720287 1.7328316
distance       0.9910789 0.9890328 0.9930971
education      1.0433604 1.0239917 1.0632190
associationyes 0.8831149 0.7594548 1.0269476
$RR
               RR       lower.ci   upper.ci
(Intercept)    0.9327691 0.8479046 1.0162244
arsenic        1.1883823 1.1577152 1.2190167
distance       0.9961905 0.9953112 0.9970557
education      1.0179727 1.0100538 1.0259153
associationyes 0.9467642 0.8813993 1.0112735
$P0
[1] 0.5751656
```

y la función R para convertir una odds ratio única en un valor de riesgo relativo:

```
# prop table
wells.tab <- with(Wells, table(switch,association))
wells.tab.p <- prop.table(wells.tab)
wells.tab
wells.tab.p
# identical
wells.or <- wells.glm.OR.RR$OR[5,"OR"]
wells.p0 <- well.glm.RR$P0
# from odds_to_rr.v2()
rr1 <- wells.glm.OR.RR$RR["associationyes","RR"]
# from or_to_rr()
rr2 <- or_to_rr( or=wells.or, p0=sum(wells.tab.p[2,]) )
```

Por razones formales, comprobamos si los resultados son idénticos.

```
> # the same?
> well.glm.RR$PO == sum(wells.tab.p[2,])
[1] TRUE
> all.equal(rr1,rr2,rr3,rr4)
[1] TRUE
```

Con el aumento de la potencia de cálculo, se impone el paradigma de los estudios de evaluación a gran escala (*large scale assessment; LSA)*, sobre todo en psicología y pedagogía empírica, que se caracterizan especialmente por muestras enormes, lo que ya justifica una referencia general a los problemas de significación. A menudo se tiene la impresión de que *cuanto más, mejor*, y esto puede ir en detrimento de un buen diseño experimental con tamaños de muestra menores. En el ámbito de la educación en particular, parece que se mide más de lo que se interviene, de modo que en psicología de la educación o en investigación educativa empírica puede aplicarse el dicho "Pesando un cerdo no se engorda", que en realidad procede de la agricultura ha llegado a la psicología de la educación y a la investigación educativa empírica. Hay indicios claros de que incluso puede ocurrir lo contrario. Medida y objeto interactúan entre sí. El fenómeno también es conocido en el sector sanitario. Wu (2009) ofrece una visión general de los problemas asociados a los estudios de LSA.

Aquí, una única prueba – operacionalizada a través del tamaño del efecto – ya puede ser innovadora en un caso individual, por ejemplo, si un enfermo terminal replicable puede curarse mediante un determinado procedimiento. Esto requiere una reconstrucción teórica y metodológicamente precisa de un efecto individual. No se necesitan necesariamente estadísticas para demostrar algo fundamental y, por tanto, un *efecto significativo*. Más bien, un buen diseño y teoría, así como una aplicación limpia, resultan ser puntos clave centrales. Por ejemplo, el representante más importante del conductismo, Burrhus F. Skinner (1904-1990), prefería trabajar muy cuidadosamente con unas pocas palomas antes que hacer estadística con grandes muestras. El trabajo de Skinner es muy preciso, brillante desde el punto de vista experimental y su obra muestra curvas de progresión (curvas de aprendizaje), etc. Los hallazgos que elaboró y la teoría del aprendizaje asociada siguen siendo el tema de todos los libros de texto de psicología básica hoy en día. Skinner está considerado uno de los psicólogos más influyentes del siglo XX, junto con Jean Piaget, Sigmund Freud y Albert Bandura (Haggbloom et al., 2002). Su precisión metodológica fue extraordinaria y está tan bien documentada que puede *replicarse*. Lo mismo puede decirse de las curvas de retención de la memoria de Hermann Ebbinghaus (1850-1909), que han sido replicadas con éxito (Murre & Dros, 2015) y más o menos en la forma que el propio Ebbinghaus postuló y antes de que se dispusiera de estadísticas exhaustivas a través de ordenadores. Así que en lugar de confiar ciegamente en las grandes muestras, vale la pena pensar en experimentos limpios y bien controlados. Esa sería una alternativa para las propias ideas de investigación. Skinner es aquí un excelente modelo a seguir.

Debemos recordar siempre que las estadísticas no lo son todo. Nuestro objetivo deben ser los *grandes efectos* y los *grandes potencias de efecto*, *no las grandes muestras*. Obviamente, una cosa tiene poco que ver con la otra. Todas las teorías más influyentes de la psicología se basan en muestras más bien modestas. Por ejemplo, el psicoanálisis y la psicología profunda, las teorías del aprendizaje, la memoria, el desarrollo cognitivo y moral, etc. Se trata de grandes diferencias que realmente marcan una diferencia en la realidad y no sólo de probar pequeñas y sutiles diferencias que pueden no tener consecuencias en la realidad.

A la inversa, también puede ocurrir que diferencias mínimas y demostrables puedan ampliarse acumulativamente hasta convertirse en grandes cambios, como la cuestión de si las personas pueden cambiarse a sí mismas y a su personalidad (Weinberger, 1994; Heatherton y Weinberger, 1994). "De Saulo a Pablo" no se da en la práctica y, desde luego, no de la noche a la mañana. Sin embargo, se producen cambios sorprendentemente profundos en las personas, de modo que se puede hablar realmente de cambios de personalidad a largo plazo, como podría demostrar un estudio de psicología profunda a largo plazo (Rudolf et al., 2001; Grande et al., 2006; Jakobson et al., 2007). Esto se corresponde con grandes potencias de efecto, que, sin embargo, no se desarrollan de la noche a la mañana, sino que son el resultado de años de trabajo (por ejemplo, Studer, 1998). El estudio sobre el éxito de la terapia catamnésica en la terapia de la adicción citado como estudio de caso (véase el capítulo 5.5.7, también Studer, 1998) es un ejemplo de ello. Todos los pasos intermedios pueden ser muy pequeños y apenas perceptibles. Experimentalmente, posiblemente

no serían ni accesibles ni detectables. Sin embargo, la diferencia durante un largo periodo de tiempo muestra cambios espectaculares, por ejemplo, cuando alguien se libra de su adicción a las drogas de muchos años y lleva una vida cambiada y más sana (Gürtler, Studer & Scholz, 2012). En otras palabras, los efectos no deben investigarse ni notificarse sin el periodo de tiempo realista subyacente. Los efectos siempre se localizan temporalmente. Por lo tanto, los efectos pueden simplemente desaparecer de nuevo sin intervención. Esto suele denominarse recuperación espontánea, un término trivial y no especialmente útil. En la literatura se conoce la *regla general de un tercio*, que es 1/3 empeorado, 1/3 igual, 1/3 mejorado. Entonces es interesante ver qué pacientes entran en cada categoría y por qué razones, y si esto se puede predecir.

A veces los efectos sólo se hacen visibles al cabo de mucho tiempo, y si el intervalo de tiempo se mide incorrectamente, los estudios comunicarían resultados sesgados e incompletos. No se trata sólo de la dirección y magnitud de los efectos, sino también de su *dinámica*, su *ritmo* y su *curso temporal*. Hay cambios que a primera vista parece que todo empeora, quizás porque las personas se enfrentan a sus crisis más profundas a través de una terapia exitosa, a la que no pueden acceder de otro modo. Si un estudio se detuviera en este punto, podría llegar a resultados muy diferentes que otro que dejara de recoger datos uno, dos o más años después. Especialmente en entornos reales y en contextos de desarrollo (educación y escuela, psicología clínica, economía) cabe esperar este tipo de patrones de cambio. Esto es contrario a las expectativas en un entorno experimental de laboratorio estrechamente definido, ya que, por definición, se adaptan al periodo de estudio y se espera que los efectos sean inmediatos. Pero, ¿pueden ser ya tan grandes, aunque sean detectables? No en vano, Meehl (1967, véase el capítulo 4.4.14.3) señala que en entornos reales la hipótesis nula es prácticamente siempre falsa, a diferencia de lo que ocurre en entornos experimentales. En la realidad, sin embargo, los tamaños de efecto deben fijarse más altos en términos de contenido, en vista de las diversas influencias contra las que tiene que afirmarse un efecto. Es cuestionable si la *d* de Cohen sigue siendo apropiada. Tendría más sentido – aunque sólo es realizable de forma incompleta – investigar contra qué y cuántas influencias y efectos en competencia debe imponerse un efecto científicamente interesante en la práctica diaria para poder tener un efecto a largo plazo. Esto da lugar a una comprensión fundamentalmente distinta de los efectos, de modo que ya no es una medida abstracta la que sirve de punto de referencia, sino la escala original y la referencia a otras variables de efecto, que lamentablemente sólo pueden reconstruirse de forma inadecuada en general.

Estos ejemplos deberían dejar claro que la estadística debe estar siempre inserta en un contexto de investigación más amplio, que debe tener una clara referencia a la realidad. Existen muchos procesos de traducción entre estos campos (Gigerenzer, 1981; Gürtler, 2005): de un lado a otro, y vice versa. Pueden producirse errores de traducción entre la teoría, el empirismo, el análisis de datos, etc., pero también pueden corregirse si las conclusiones se abren en abanico siguiendo diversos criterios y no sólo basándose en unos pocos indicadores.

En definitiva, por tanto, la idea expresada en varias ocasiones nos parece que tiene mucho sentido para seleccionar modelos complejos, integrar casos especiales junto al caso general, utilizar en la medida de lo posible los conocimientos previos y emplear métodos gráficos de comprobación de modelos. A ello se añaden simulaciones y predicciones basadas en los modelos estimados para compararlos con los datos empíricos o con nuevos datos. El resultado es una gran cantidad de información que, en conjunto, debería permitir una evaluación seria de la situación de los datos en comparación con las hipótesis y los modelos teóricos. De este modo, el análisis de los datos pasa de la comprobación a la estimación (contra-argumentos, por ejemplo, Mayo, 2018). En última instancia, todos los resultados numéricos deben probarse cualitativamente si tienen sentido y se ajustan a la teoría o la contradicen. Pueden producirse paradojas; la paradoja de *Simpson*, la de *Lindley-Jeffreys* y la de *Meehl* son contradicciones especialmente interesantes.

4.4.14 Paradojas en estadística

4.4.14.1 Paradoja de Simpson

Esta paradoja según Edward H. Simpson (1922-) muestra que el análisis de un grupo entero puede conducir a resultados completamente diferentes que el análisis por separado de los subgrupos individuales que

componen el grupo entero. El efecto se produce sobre todo en tablas de cuatro campos (Simpson, 1951; Kievit, Frankenhuis,Waldrop & Borsboom, 2013). El efecto se dio a conocer a través de la acusación contra la Universidad de California/ Berkeley de que las mujeres tenían menos posibilidades de obtener una plaza de estudios de posgrado en su solicitud que los hombres (Bickel, Hammel & O'Connell, 1975). El análisis global resultó ser estadísticamente significativo desde el punto de vista clásico. El script R contiene algunos análisis y tablas más; nos limitamos a lo esencial (`ptII_quan_classicstats_Simpsonparadox.r`) y lo representamos como un gráfico de asociación (véase la Fig. 4.67):

```
# assoc plot
# that almost shows everything ...
assoc(aperm(UCBAdmissions), expected = ~ (Admit + Gender) * Dept,
compress = FALSE,
labeling_args = list(abbreviate = c(Gender = TRUE),
rot_labels = 0))
```

Y así, al principio, supuestamente parece discriminación:

```
> # admission versus rejected for male versus female
> ucbd.tab <- margin.table(UCBAdmissions, margin=c(1,2))
> ucbd.tab
Gender
Admit Male Female
Admitted 1198 557
Rejected 1493 1278
> chisq.test(ucbd.tab)
Pearson's Chi-squared test with Yates' continuity correction
data: ucbd.tab
X-squared = 91.61, df = 1, p-value < 2.2e-16
```

Sin embargo, un análisis más detallado de las distintas facultades y departamentos demostró que las mujeres no estaban sistemáticamente en desventaja. Mostramos aquí un gráfico de mosaico (véase la Fig. 4.68)

```
mosaicplot(UCBAdmissions, border="darkblue", shade=TRUE)
```

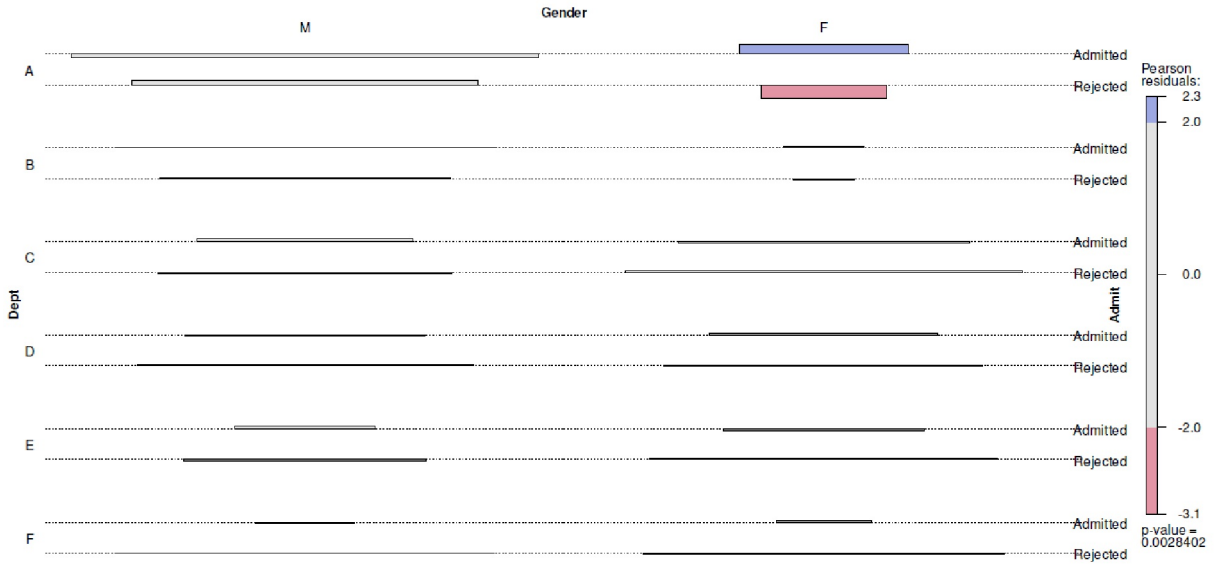


Figura 4.67. *Paradoja de Simpson (Gráfico de asociación)*

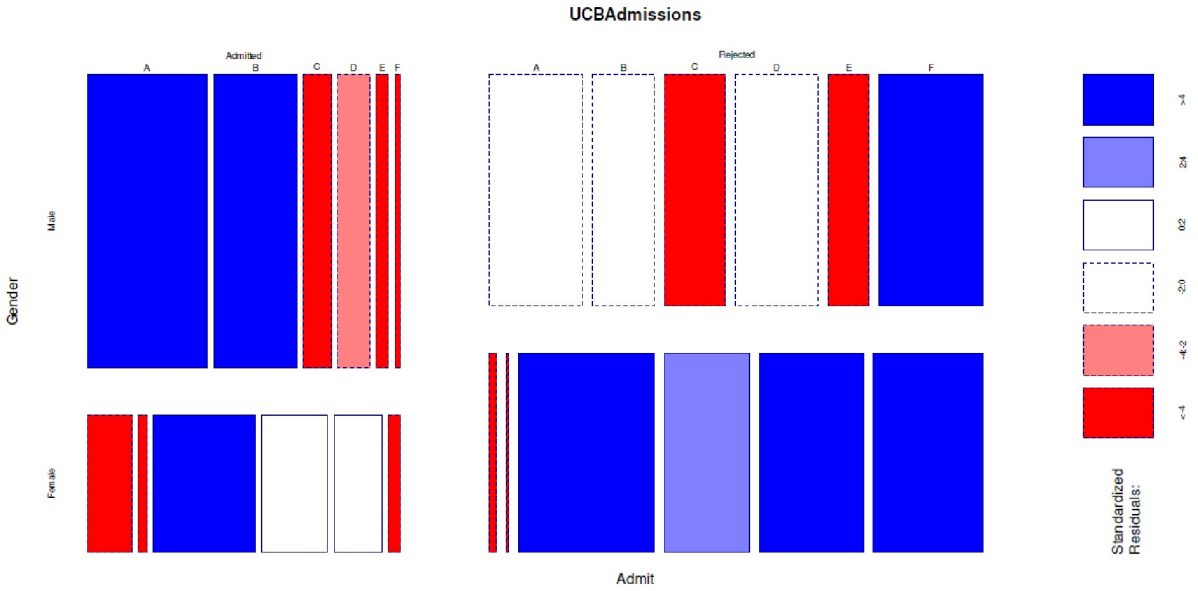


Figura 4.68. *Paradoja de Simpson (Gráfico de mosaico)*

Sin embargo, se observó que las solicitudes de hombres y mujeres no se distribuían por igual entre los departamentos (una diferencia que era igualmente significativa desde el punto de vista estadístico). Un análisis más detallado reveló que las mujeres solicitaban plaza allí donde había menores tasas globales de admisión para ambos sexos, mientras que los hombres lo hacían allí donde, en general, había mayores tasas de admisión. La conclusión es que las mujeres no estaban en desventaja en este caso, sino que en realidad tenían una pequeña ventaja. Esto no significa que en otros lugares no se discrimine sistemáticamente a las personas en función de su sexo, edad, origen, etc., que así sean (o puedan ser) sistemáticamente discriminadas. El análisis detallado tiene en cuenta esta información. La lección que cabe extraer de ello es que las afirmaciones globales sólo están fiables hasta cierto punto y es mejor interpretarlas en el contexto de subgrupos individuales. En particular, deben tenerse en cuenta las tasas de base (Gigerenzer, 1991; Gigerenzer & Horage, 1995; Krynski & Tenenbaum, 2007) y no sólo las afirmaciones porcentuales globales. De lo contrario, se produce una confusión entre estadística y lógica proposicional (Pearl, 1999-04). Pearl (2009, p.78) llega a decir que "cualquier relación estadística entre dos variables puede invertirse cuando se incluye factores adicionales en el análisis."

Lindley y Novick (1981) formulan los principales rasgos de la paradoja de Simpson del siguiente modo: ninguna estadística protege a los investigadores de extraer conclusiones sustancialmente incorrectas. Las estadísticas tampoco hacen afirmaciones sobre qué valores numéricos son correctos y cuáles incorrectos. La paradoja de Simpson afecta tanto a la estrategia de muestreo, la estrategia de evaluación asociada (es decir, según qué características y aspectos de la muestra se evalúan) como al diseño básico o la cuestión de la abstracción, es decir, a qué nivel y con qué grado de detalle deben extraerse las conclusiones. La paradoja de Simpson se produce, como ya se ha explicado en el caso de la UC Berkeley, cuando los índices de base de los distintos subgrupos de una muestra son diferentes y la agregación a nivel abstracto da lugar a categorías que ya no reflejan adecuadamente estas diferencias. En el ejemplo, estas son las diferentes tasas de candidatos por sexo para los distintos departamentos. En R, (`ptII_quan_classicstats_Simpsonparadox.r`) se puede simular la paradoja de Simpson fácilmente:

```
# Simpson Paradox
# simulation
seed <- 99883
set.seed(seed)
# create some artifical data that represent the Simpson Paradox
v <- data.frame(x=c(c(1:10)+rnorm(10), c(11:20)+rnorm(10)),
y=c(c(10:1)+rnorm(10), c(20:11)+rnorm(10)), group=gl(2,10))
# plot two groups
```

```
SP.sim(v)
```

La salida de SP.sim(v) es la siguiente: una estimación global y la de los subgrupos (véase la Fig. 4.69):

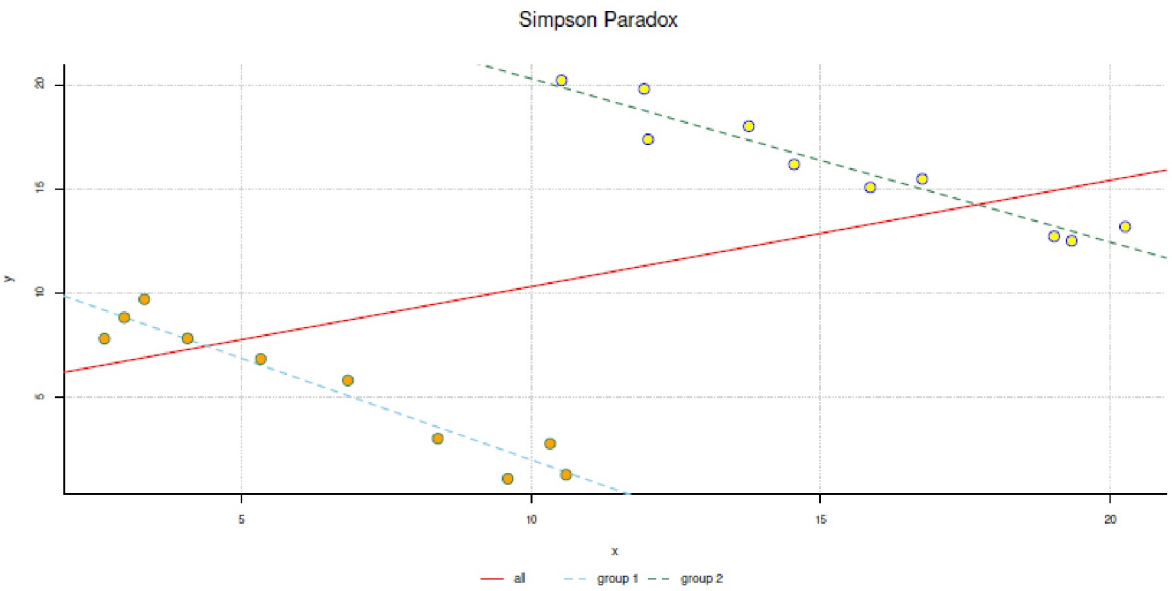


Figura 4.69. *Paradoja de Simpson (Simulación)*

```
lm(formula = y ~ x, data = v)
coef.est coef.se
(Intercept) 5.22 2.79
x 0.51 0.23
---
n = 20, k = 2
residual sd = 5.60, R-Squared = 0.22
lm(formula = y ~ x, data = subset(v, group == 1))
coef.est coef.se
(Intercept) 11.76 0.71
x -0.98 0.10
---
n = 10, k = 2
residual sd = 0.94, R-Squared = 0.92
lm(formula = y ~ x, data = subset(v, group == 2))
coef.est coef.se
(Intercept) 28.16 1.29
x -0.78 0.08
---
n = 10, k = 2
residual sd = 0.84, R-Squared = 0.92
```

El paquete `Simpsons` de R con la función `Simpsons()` intenta encontrar conglomerados en datos bivariantes continuos y asegurarlos mediante pruebas de permutación y regresión para encontrar subgrupos.

```
# detect subgroups
out.s <- Simpsons(x,y,data=v)
summary(out.s)
# check clustering
print(out.s)
coef(out.s)
```

Abreviado, se pueden observar los coeficientes de la estimación así como comparar la agrupación original con la estimación mediante `mclust`:

```
# check clustering
coef(out.s)
str(out.s)
out.tab <- cbind(v, mclust=out.s$mclustanalysis$classification)
cbind(out.tab, comp=out.tab[,"group"] == out.tab[,"mclust"])
```

La paradoja de Simpson demuestra, sobre todo, que los análisis nunca son definitivos, ya que nunca se pueden descifrar todas las influencias relevantes. Y muestra la necesidad de modelos complejos tanto para identificar como para modelizar los subgrupos relevantes, para llegar a conclusiones más válidas.

4.4.14.2 La paradoja de Jeffreys-Lindley

La paradoja Jeffreys-Lindley fue descrita por primera vez por Sir Harold Jeffreys (1939/1961) y posteriormente denominada paradoja por Lindley (1957) y desarrollada por Bartlett (1957). La paradoja de Jeffreys-Lindley es una situación en la que los enfoques frecuentista y bayesiano llegan a conclusiones diametralmente opuestas sobre una prueba de hipótesis debido a sus respectivas especificidades. En este caso, la hipótesis nula H_0 puede probarse simultáneamente

- según el *enfoque frecuentista* y se puede rechazarla al nivel clásico del 95% (= rechazo de H_0) y al mismo tiempo
- según el *enfoque bayesiono* y se puede asignár una probabilidad posterior del 95% (= aceptación de H_0).

Así pues, existen indicaciones frecuentistas legítimas para el rechazo de H_0, así como indicaciones bayesianas para la aceptación de H_0, cada una considerada desde su propia lógica. Es una paradoja en la medida en que va en contra de las expectativas, pero en la discusión dialéctica conduce a una mejor comprensión de la teoría estadística respectiva. Una vez más, no se trata de una paradoja, ya que es simplemente la aplicación coherente de la teoría estadística respectiva, cuyas diferencias, sin embargo, sólo se hacen tan evidentes en determinadas condiciones.

Para ilustrar el fenómeno, se reproduce a continuación un ejemplo numérico y ampliamente utilizado (Wikipedia, 2020) sobre el tema. Es similar al ejemplo de Laplace sobre la natalidad de las niñas con datos recogidos en París entre 1745 y 1770. El punto de partida ficticio es ahora que en una determinada ciudad en un determinado periodo de tiempo nacen 49.581 niños y 48.870 niñas. La proporción del número de nacimientos es, por tanto, 49 581/48 870 $\approx$ 1.012 a favor de los niños. La proporción respectiva es 50.36% para los niños y 49.64% para las niñas. La pregunta que se plantea ahora es si el parámetro θ de la distribución binomial del número de nacimientos para niños y niñas es $\theta = 0.5$ o un valor diferente. Las dos hipótesis en liza son las siguientes

$H_0 : \theta = 0.5$
$H_1 : \theta \neq 0.5$

El enfoque frecuentista se puede calcular con la prueba binomial `binom.test()` en R o a mano. En vista de la gran muestra con $n = 49\,581 + 48\,870 = 98\,451$ una distribución normal puede actuar como solución aproximada en lugar de la distribución binomial con $X \sim N(\mu; \sigma^2)$, donde $\mu = n\theta$ y $\sigma = n\theta(1 - \theta)$. θ denota la probabilidad binomial p. Se realiza una prueba bilateral de modo que se tiene doblar el valor p unilateral calculado manualmente. Para la prueba se toma el número k de chicas (`ptII_quan_classicstats_-JeffreysLindleyparadox.r`).

```
> # Wikipedia data
> # https://en.wikipedia.org/wiki/Lindley%27s_paradox#Numerical_example
> # observed data
> gb <- c(boys=49581, girls=48870)
> gb
boys  girls
49581 48870
```

```
> # sum
> n <- sum(gb)
> n
[1] 98451
> # observed proportions
> gb/n
boys       girls
0.5036109 0.4963891
> # frequentist solution
> # H0: theta = 0.5
> # H1: theta != 0.5
> theta <- 0.5
> # normal approximation
> # X ~ N(mu, sigma^2)
> # mu = n*p = n*theta
> mu <- n * theta
> mu
[1] 49225.5
> # var = #= n * theta *(1-theta)
> sigma2 <- mu*(1-theta)
> sigma2
[1] 24612.75
> # p-value
> # binom.test(n-k,n,theta)
> k <- gb["girls"]
> k <- gb["boys"]
> binom.test(n-k,n,theta)
Exact binomial test
data: n - k and n
number of successes = 48870, number of trials = 98451, p-value = 0.02365
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
0.4932610 0.4995174
sample estimates:
probability of success
0.4963891
> ND.fun <- function(u) 1 / sqrt(2*pi*sigma2) *
+          exp(-((u-mu)^2)/(2*sigma2))
> coveredarea <- integrate(ND.fun, lower=k, upper=n)
> # two-sided
> pv1 <- 2*(coveredarea$value)
> pv1
[1] 0.0234515
> #= normal approximation
> pv2 <- pnorm(k,n,sigma2)
> pv2
boys
0.02354131
> # check
> all.equal(pv1,pv2)
[1] "names for current but not for target"
[2] "Mean relative difference: 0.003829461"
> abs(pv1-pv2)
boys
8.980662e-05
```

Como puede observarse, el valor *p* es inferior al nivel de convención del 5%. Así pues, la visión frecuentista rechaza la H_0 a un nivel convencional del 5%. Por lo tanto, no hay que suponer una distribución igual del número de nacimientos entre los sexos, sino rechazarla.

El resultado bayesiano es el siguiente: partiendo del teorema de Bayes (véase cap. 6.4) para el caso discreto, primero se formula la expectativa a priori, que consiste en que no hay pruebas para asignar una probabilidad a priori mayor a una de las dos hipótesis que a la otra. Por tanto, la elección es $p(H_0) = p(H_1) = 0.5$. Además, se supone una distribución uniforme de θ bajo H_1. Ahora se pueden calcular los elementos

del teorema de Bayes. Por *n* denotamos el número total de nacimientos y por *k* el número de nacimientos de interés, aquí k = 48 870 para el número de niñas nacidas.

```
> # Bayes solution
> # prior
> # equal for H0 and H1
> prior.H0 <- 0.5
> prior.H1 <- 0.5
> # Likelihood H0
> k <- gb["boys"]
> p.k.H0 <- exp(lchoose(n,k) + log(theta)*(k) + log(1-theta)*(n-k))
> p.k.H0
boys
0.000195
> # Likelihood H1
> k <- gb["girls"]
> p.k.H1 <- exp(lchoose(n,k) + lbeta(k+1,n-k+1))
> p.k.H1
girls
1.02e-05
```

Para una comprensión más sencilla, se pueden introducir estrictamente los datos en el teorema de Bayes:

```
> # applying Bayes' Theorem
> # p(H0|k)
> p.H0.k <- p.k.H0 * prior.H0 / (p.k.H0*prior.H0 + p.k.H1*prior.H1)
> # p(H1|k)
> p.H1.k <- p.k.H1 * prior.H1 / (p.k.H0*prior.H0 + p.k.H1*prior.H1)
> # prob in favor of H0
> p.H0.k
boys
0.950523
> # prob in favor of H1
> p.H1.k
boys
0.04947704
> 1-p.H0.k
boys
0.04947704
> # odds ratio of H0/H1 in favor of H0
> p.H0.k/p.H1.k
boys
19.21139
```

Esto da como resultado una probabilidad posterior de $p(H_0|k) = 0.951$ y $1 - p(H_0|k) = p(H_1|k) = 0.0495$, de modo que desde el punto de vista bayesiano se puede suponer – y no sólo mantener – H_0. Se trata de una circunstancia que no sería posible en la estadística frecuentista. La odds ratio posterior H_0/H_1 es de 19.2, lo que habla claramente en favor de H_0. Esto se puede expresar gráficamente (véase la Fig. 4.70):

```
# plot H0 with k=n/2 i.e. p=0.5 against p=k/n
sek <- seq(0.48,0.52,length.out=1000)
p.k.H0.fun <- function(theta, n, k) exp(lchoose(n,k) +
           log(theta)*(k) + log(1-theta)*(n-k))
p.k.H0.sek <- p.k.H0.fun(sek, n=n, k=n/2)
p.k.H1.sek <- p.k.H1.brob.fun2(sek)
plot(sek,p.k.H0.sek, col="darkred", xlab=expression(theta),
           ylab="p(k)", bty="n", pre.plot=grid(), type="l",
main="Jefferys-Lindley Paradox")
lines(sek, p.k.H1.sek, col="green")
legend("topright", legend=c("p(k|H0) with k=n/2 (p=0.5)",
           "p(k|H1) with k=empirical data"), col=c("darkred","green"),
           lty=c(1,1), bty="n")
```

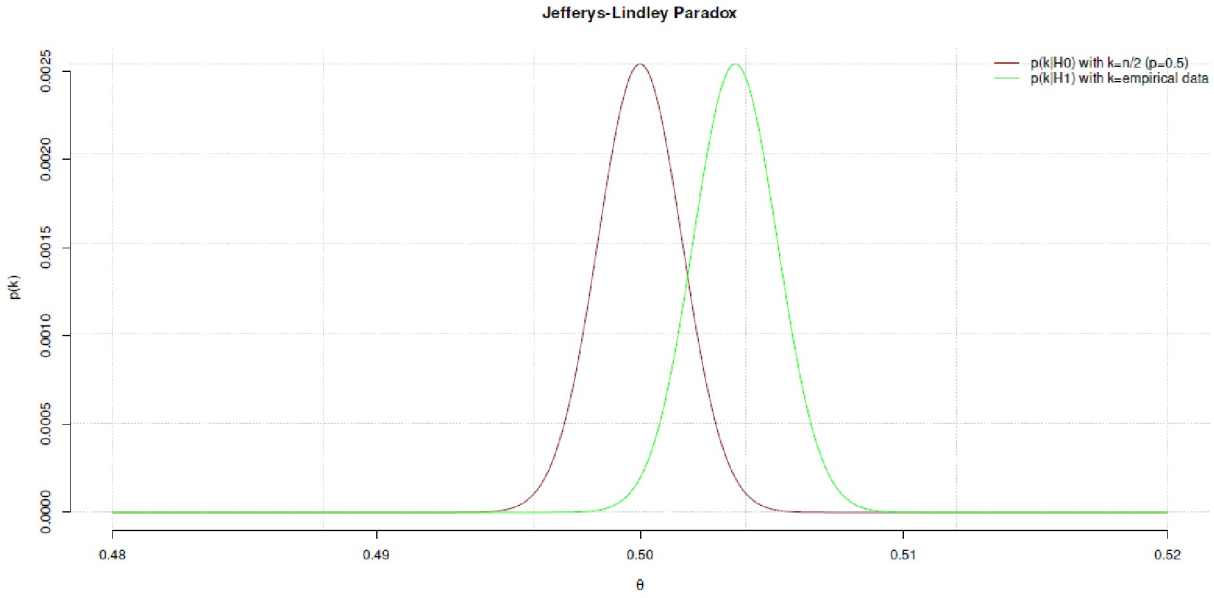


Figura 4.70. *Jeffreys-Lindley Paradox (distribución posterior)*

Esta situación fundamental ha dado lugar a una plétora de artículos y el interés por esta paradoja continúa hasta nuestros días (entre otros, Sprenger, 2012; Spanos, 2013; Robert, 2013; Villa & Walker, 2015-03-13). El fenómeno puede encontrarse en los siguientes marcos, en los que el enfoque difiere según el artículo y el autor, y en algunos casos pueden encontrarse afirmaciones contrarias. En este sentido, las siguientes afirmaciones son más bien una descripción aproximada de las condiciones en discusión y se requiere una aclaración concreta de las respectivas cadenas de argumentación:

- La lógica proposicional de la prueba difiere fundamentalmente entre la bayesiana y la frecuentista:

  - La prueba de la hipótesis nula H_0 (frecuentista) no tiene en cuenta H_1, mientras que con Bayes se contrasta explícitamente H_0 con H_1.
  - El frecuentista encuentra que H_0 no proporciona una buena explicación de los datos observados y, por lo tanto, la rechaza en el nivel de exceso crítico.
  - El bayesiano prueba el poder explicativo relativo de H_0 frente a H_1 y encuentra que H_0 en relación con H_1 explica mejor los datos. Esto no excluye la posibilidad de que existan H_2, H_3, etc. que puedan explicar los datos incluso mejor que H_0.

- En el ejemplo bayesiano, la H_1 se formula de forma mucho más difusa que la H_0, ya que se supone que x es igual en todo el espectro de [0; 1]. En cuanto a la H_0, se examina específicamente x = 0:5, a la vista de las observaciones realizadas. La H_1 pregunta lo mismo, pero no sólo concretamente para $\theta = 0.5$ (= hipótesis puntual), sino para todo el espectro [0; 1] (= hipótesis difusa). Se trata de comparar una hipótesis clara frente a otra muy poco clara. Esto tiene un efecto en lo siguiente. Cambiar la probabilidad a priori para H_1 cambiaría la situación y podría conducir a un resultado diferente de la prueba. A este respecto, Christensen (2010-07-16, p.1) dice: "La moraleja de la paradoja de Lindley-Jeffreys es que si eliges una probabilidad a priori estúpida, puedes obtener una probabilidad a posteriori estúpida." De ello se sigue, para bajo la H_1, elegir una distribución adecuada y no necesariamente una que consista en ignorancia (= distribución a priori no informativa) y afirme que "podría ser cualquier cosa" – incluso o especialmente porque en la mayoría de los contextos debería ser posible derivar hipótesis razonables para la H_1 a partir del conocimiento contextual disponible. En nuestro ejemplo de la distribución por sexos y la tasa de

natalidad, el sentido común no pensaría en asumir a priori 8:2, 9:1 u otras proporciones extremas, por ejemplo. Tales valores caerían en el ámbito de la H_1 difusa. Quienes formulan a priori sin ninguna razón no deberían sorprenderse si no sale nada de ello. En una entrada de blog, Gelman (2013b) comenta lo siguiente:

> „To me, the Lindley paradox falls apart because of its noninformative prior distribution on the parameter of interest. If you really think there's a high probability the parameter is nearly exactly zero, I don't see the point of the model saying that you have no prior information at all on the parameter. In short: my criticism of so-called Bayesian hypothesis testing is that it's insufficiently Bayesian. [...] I'm speaking of all the examples I've ever worked on in social and environmental science, where in some settings I can imagine a parameter being very close to zero and in other settings I can imagine a parameter taking on just about any value in a wide range, but where I've never seen an example where a parameter could be either right at zero or taking on any possible value. But such examples might occur in areas of application that I haven't worked on."

Bartlett (1957) hace una afirmación comparable, a saber, que los factores de Bayes (véase el capítulo 6.8.1.4) reaccionan sensiblemente a la amplitud de una distribución uniforme. Así, un cambio en la distribución provoca cambios en los factores de Bayes como una actualización de las expectativas.

- Desde un punto de vista frecuentista, se nota una muestra muy grande y se aplican las observaciones sobre la dependencia de los errores estándar del tamaño de la muestra o de la significación. No es sorprendente que se encuentre un efecto potencialmente pequeño con una muestra muy grande. Ahora habría que preguntarse si este efecto tiene un tamaño sustancial que implique consecuencias en la realidad. La paradoja de Jeffreys-Lindley no especifica el tamaño del efecto. Pero en términos del tamaño de la muestra tiene el efecto de que si el rango de parámetros para θ y el umbral crítico α se mantienen constantes, al aumentar el tamaño de la muestra n la H_0 se rechaza simultáneamente con $p < \alpha$ y el factor de Bayes contra la H_0 es $BF < \alpha$. Según Wagenmakers (en Gelman, 2013b) basta con que el umbral crítico α se mantenga constante y $n \rightarrow \infty$ para que se produzca este fenómeno. Según el autor, la distribución a priori es menos relevante. Sin embargo, se produce una situación paradójica si (Hubbard & Lindsay, 2008, p.76, cursiva en el original) "para *n* grande, un valor de *p* pequeño puede interpretarse en realidad como una prueba *a favor de* H_0 en lugar de *en su contra*. La cuestión de la objetividad y la utilidad del valor *p* como medida de la evidencia se hace añicos con este argumento". Según LaMont y Wiggins (2016-10), el fenómeno se produce (ibíd., p.1) "cuando las hipótesis en competencia (modelos estadísticos) tienen dimensiones diferentes. En este contexto, la inferencia bayesiana depende sensiblemente de la elección de la distribución a priori [...] y el nivel de confianza frecuentista implícito suele superar con creces el 95 %". Según los autores, la explicación matemática se sitúa en el contexto de la navaja de Occam, es decir, la cuestión de la eficacia y la simplicidad (ibíd., p.4). Esto se operativiza como la probabilidad de extraer del espacio de parámetros de la distribución a priori de forma aleatoria y coherente a lo largo de los supuestos del modelo. En la estadística bayesiana, los parámetros se consideran variables aleatorias y no como variables fijas. En consecuencia, cada parámetro tiene su propia distribución. En la aplicación, se deduce que (ibíd., cursiva en el original)

  > „Complex models (large K) with uninformative priors have small Occam Factors, due to the large volume of plausible parameters (V_0), relative to the volume of parameter space consistent with the observations (V_N). Therefore the Occam factor automatically penalizes complex models and is a natural mathematical realization of the Occam Razor: *Among competing hypotheses, the one with the fewest assumptions [parameters] should be selected.*"

  En el caso de la paradoja de Bartlett (a veces tratada como una variante de la paradoja de Jeffreys-Lindley en la bibliografía), la probabilidad de extraer un parámetro coherente con el modelo disminuye hacia cero, mientras que el espacio de parámetros posibles $V_0 \rightarrow \infty$. La paradoja de Jeffreys-Lindley es, según los autores, una generalización de esta situación para V_0 finito. La probabilidad a priori de los distintos parámetros posibles – en contraste con el argumento anterior

> de Wagenmaker, que considera que el papel de la probabilidad a priori es más bien modesto – no sólo determina la probabilidad a priori, sino que "también establece *la escala* de un tamaño de efecto 'típico'" (ibíd.). Una consecuencia importante de esto es que la *prior* puede ser poco informativa con respecto a las estimaciones de los parámetros, pero no lo es con respecto al propio modelo. En este caso, la a priori es *muy informativa por su propia forma, y esto tiene consecuencias*.

Los comentarios anteriores sólo reflejan una parte muy pequeña de los argumentos matemáticos y estadísticos que se pueden encontrar sobre este tema. Independientemente de los detalles y significados matemáticos podemos aprender de ello que la contextualidad de los resultados y la influencia del conocimiento previo, así como la elección de los parámetros, no deben subestimarse, ya que tienen un impacto duradero en el enfoque, la comprensión del problema, el plan de estudio y la perspectiva. Aunque, desde una perspectiva frecuentista, el conocimiento previo no debería influir en las ecuaciones, lo hace de forma impresionante y persistente de muy diversas maneras, ya sea explícitamente como en el caso de Bayes o implícitamente como en el de los frecuentistas. Esta circunstancia – ignorar y ocultar el conocimiento contextual – es uno de los mayores problemas de la estadística frecuentista. Por supuesto, la información contextual se incluye, ya sea como análisis de potencia a priori en el diseño de un estudio o de otras formas, como los múltiples supuestos sobre distribuciones, des-/dependencias y hetero- u homogeneidades (por ejemplo, de varianzas), los supuestos sobre niveles de significación, tamaños de efecto significativos, etc. Todos ellos son supuestos que, por ejemplo, deben tenerse en cuenta en el diseño de un estudio. Todos estos son supuestos que en algunos casos ni siquiera se originan en el contexto, lo que empeora aún más la situación. Por tanto, parece sensato controlar metódicamente los conocimientos previos, utilizarlos de forma consciente y directa y no pretender que los científicos sean ingenuos. Esto debe distinguirse del hecho de que, por ejemplo, en el análisis secuencial cualitativo (véase el capítulo 11.9 sobre el procedimiento) el conocimiento contextual se omite temporal pero intencionadamente para poder formular hipótesis de base amplia y no dejarse llevar por presuposiciones erróneas. La planificación de un diseño cuantitativo requiere un enfoque similar para no caer presa de los propios supuestos y utilizar la información de forma que no se limite innecesariamente el espacio de hipótesis de antemano.

La paradoja de Jeffreys-Lindley es un excelente ejemplo de cómo la inclusión y alteración de información contextual puede cambiar los resultados estadísticos o nueva información puede cuestionar un resultado supuestamente claro. Otra conclusión sería que el mundo, como producto de nuestras construcciones, se comporta exactamente igual que la información que se interpreta en él. Como resultado, cambian las conclusiones sobre las decisiones o los hallazgos que hay que tomar o los conocimientos que hay que adquirir. Desde un punto de vista informativo-teórico y de causa-efecto esto es inmediatamente comprensible: la observación interactúa de muchas maneras con la recogida y el análisis de datos.

Pasemos ahora a la conexión entre la estadística y las buenas teorías, lo que es el contenido de la paradoja de Meehl.

4.4.14.3 La paradoja de Meehl

Paul Meehl (1920-2003), antiguo presidente de la Asociación Americana de Psicología (APA), describe en su influyente artículo (Meehl, 1967) una situación de investigación bastante común en la que el aumento de la precisión experimental mediante el uso de pruebas de significación de hipótesis nulas (véase Null-Ritual, Gigerenzer, 2004b; Gigerenzer, Krauss & Vitouch, 2004) conduce a una disminución de la calidad de las teorías. Esto es muy perjudicial para la comprobación de las teorías en el sentido popperiano, ya que parece que se falsifican con demasiada rapidez y facilidad o que se aceptan cosas nuevas con demasiada rapidez. En ambos casos, el trabajo teórico y la planificación cuidadosa del diseño se quedan en el camino. En concreto, esto está relacionado con la ya comentada asignación u orientación sustantiva de H_0 o H_1 y H_2.

Por ejemplo, Meehl (1967) predice que en los diseños no aleatorios la probabilidad de rechazar la hipótesis nula (= sin diferencias de grupo) es aproximadamente $p = 0.5$ para una hipótesis alternativa dirigida. Para hipótesis alternativas no dirigidas, se supone una $p = 1$, es decir, ¡se encuentran diferencias en el 100% de los casos!

Meehl asume que en entornos reales – a diferencia de los experimentales – la hipótesis nula es siempre falsa. Sólo necesita una fuerza de prueba suficiente o tamaños de muestra máximos para encontrar diferencias, ya que el valor *p* como garante de la significación depende directamente del tamaño de la muestra. Gigerenzer (2004b) informa de estudios (Waller, 2004; Bakan 1966) que intentaron probar empíricamente esta predicción utilizando estudios empíricos publicados y simulaciones. En un amplio estudio con una hipótesis alternativa dirigida elegida arbitrariamente, los autores pudieron confirmar este supuesto para el 46% de las predicciones (tamaño de la muestra N = 81 000, 511 ítems del MMPI-2), en algunos casos con valores *p* impresionantemente pequeños. Los resultados llaman directamente la atención sobre las fatales consecuencias de publicar pruebas de significación con muestras grandes y *p*-valores pequeños – pero sin tamaño del efecto, potencia y teoría razonable. La paradoja descrita puede encontrarse en el ámbito de los estudios de evaluación a gran escala. Estos estudios hacen hincapié en el gran tamaño de las muestras como su característica más destacada y suelen acabar en modelos de ecuaciones estructurales que son complejos pero que, al mismo tiempo, pueden enmascarar los tamaños del efecto central subyacente y, además, prácticamente nunca se someten a una prueba crítica de lo contrario.

4.5 Conclusión: Fisher frente a Neyman-Pearson de un vistazo

Para concluir con la estadística clásica, compararemos una vez más el enfoque de Fisher con el de Neyman-Pearson en forma de tablas para acotar específicamente las diferencias y similitudes. Además, discutiremos los mitos que rodean al valor p en una tabla. La razón de ello es que los cursos de estadística y los libros de texto siguen sin explicar las teorías de Fisher y de Neyman-Pearson – se presentan adecuadamente las teorías de Fisher y Neyman-Pearson, pero se descuidan por completo los detalles técnicos, es decir, "¿Qué es un valor p?" y "¿Qué podemos decir con él?" (Gigerenzer, Krauss & Vitouch, 2004). Las excepciones son Eid, Gollwitzer & Schmitt (2010) y algunos artículos de revistas (Hubbard, 2004; Gigerenzer, 2004b) que abordan este tema en detalle. Los antecedentes históricos que consideramos relevantes también quedan fuera de los libros de texto. Se pueden encontrarlos en Jaynes (2003, cap. 16) y, en parte, en Gigerenzer y Marewski (2015).

Aunque las matemáticas de ambos enfoques son en última instancia idénticas y no se critican como tales, cada uno de los fundadores basa sus teorías en una comprensión muy diferente de la estadística: la inferencia en Fisher (véase el capítulo 4.3.2) y el comportamiento o la toma de decisiones en Neyman-Pearson (véase el capítulo 4.3.3). Así pues, los valores p (Fisher) no son tasas de error (Neyman-Pearson), como destacan Hubbard (2004) y Hubbard y Bayarri (2003). La significación es y sigue siendo una cuestión difícil (Hubbard & Ryan, 2000). Lo que se desprende de todo este trabajo es que, a pesar de utilizar los mismos procedimientos matemáticos, las diferencias cualitativas y las interpretaciones son mucho mayores de lo que puede leerse en los libros de texto actuales. Las tablas 4.14, 4.15 y 4.16comparan los valores p y los porcentajes de error de Fisher y Neyman-Pearson con respecto a los conceptos, las teorías y su significado. Termina con una lista de mitos torno al valor p en la Tabla 4.17.

Tabla 4.14: *Comparación de los conceptos Fisher frente a Neyman Pearson*

Concepto	Fisher	Neyman-Pearson
Hipótesis nula H_0 (hipótesis real NIL)	(sí)*	-
Hipótesis H_1 e hipótesis alternativa H_2	-	sí
límite de significación convencional (exceso de probabilidad crítica)	(sí)*	sí
notificar el valor p exacto	(sí)*	sí
sólo es relevante la decisión de la prueba, pero no el valor p	-	sí
conocimiento inductivo/descubrimiento de algo nuevo debido al rechazo de la hipótesis nula	sí	-
decisión inductiva basada en la prueba estadística y – en caso ideal – el análisis previo de las variables contextuales	-	sí
análisis de potencia a priori	-	sí
tamaño de la muestra N	-	sí
tasa de error α	-	sí
tasa de error β	-	sí
Potencia (fuerza de la prueba) $1 - \beta$	-	sí
significación práctica/fuerza del efecto	-	sí

*para el Fisher temprano, pero no para el Fisher tardío

Se hacen más distinciones cuando se compara directamente el Fisher temprano con el Fisher tardío. se compara:

Tabla 4.15: *Comparación de Fisher temprano y tardío frente a Neyman-Pearson*

Elementos teóricos y conceptos	Fisher ~1935	Fisher ~1955/56	Neyman-Pearson
Número de hipótesis	1	1	2
NIL hipótesis nula	sí	-	-
no necesariamente NIL hipótesis nula	-	sí	sí
(sin especificar) Hipótesis nula - Prueba de significación (NHST), es decir, ritual nulo / en Gigerenzer, Krauss y Vitouch (2004)	-	-	-
Indicación de NHST	-	-	-
Hipótesis alternativa	-	-	sí
Aceptación de la hipótesis nula/H1	-	-	sí
Aceptación de la hipótesis alternativa	-	-	sí
Rechazo de la hipótesis nula/H1	sí	sí	sí
Rechazo de la hipótesis alternativa	-	-	sí
Tamaño exacto del valor p pertinente	-	sí	-
Exceso de probabilidad crítica	Convención	p exacto	Análisis de potencia
Nivel de significación		Informe	
Convención(es) para el nivel de significación	sí		-
Informe de valores p exactos	-	sí	-
Tasa de error α / Error de tipo I	-	-	sí
Tasa de error β / Error de tipo II	-	-	sí
Error de tipo III	-	-	-
Error de tipo IV	-	-	-
Análisis del diseño según Gelman y Carlin (2014)	-	-	-
Potencia	-	-	sí
Intervalos de confianza	-	-	sí
Potencia de efecto	-	-	sí
Requisito de replicación	-	-	sí
Afirmaciones basadas en un único estudio	-	-	sí
Afirmaciones basadas en muchos estudios	-	-	sí
Uso cuando se sabe poco sobre un área de investigación	-	sí	-
Planificación previa de los parámetros del estudio	-	-	sí
Planificación del tamaño de la muestra	-	-	sí
Lógica proposicional	inductiva	inductivo o asunto des-conocido	hipotético-deductivo
Nuevos conocimientos para la ciencia	sí	sí	-
Comportamiento puro/ decidir en base a resultados	-	-	sí
p (hipótesis | datos)	-	-	-
p (datos | hipótesis)	sí	sí	sí
distribuciones posteriores (probabilidad de parámetros)	-	-	-
Inclusión de información previa	-	-	-

A continuación se compara el valor *p* (Fisher) y su utilización frente al índice de error α (Neyman-Pearson):

Tabla 4.16: *Comparación del valor p y la tasa de error α*

Valor *p*	Tasa de error α
Prueba de significación	Prueba de hipótesis
Prueba contra H_0	Tasa de error α de tipo I
Filosofía inductiva y cognición	Enfoque deductivo por diseño
Pautas inferenciales para interpretar las pruebas en los datos	Pautas de comportamiento para tomar decisiones basadas en los datos (objetivamente)
Variable aleatoria basada en datos	Valor predeterminado antes empirismo
Propiedad de los datos	Propiedad de la prueba
válida a corto plazo y aplicable a un único estudio / experimento	válida a largo plazo y aplicable a réplicas continuas e idénticas
Población hipotéticamente infinita	Población claramente delimitada y definida en el contexto del análisis de potencia a priori, pero las conclusiones se extraen, si es posible, de muestras infinitamente grandes
interpretable de forma independiente	interpretable sólo en el contexto de tasa de error β, potencia, fuerza del efecto y el tamaño de la muestra

Y ahora algunos mitos e ilusiones más en torno al valor *p* (entre otros Gigerenzer, 2004b; Hubbard, 2004; Hubbard & Lindsay, 2008):

Tabla 4.17: *Mitos sobre el valor p*

Significado del valor p	¿Es esto correcto?
indica la probabilidad de los datos empíricos o incluso más extremos bajo la validez de la hipótesis nula	sí (= denotación exacta del valor p)
indica la probabilidad de sólo los datos empíricos bajo la validez de la hipótesis nula	no, los valores p siempre incluyen la probabilidad de datos aún más extremos bajo la validez de la hipótesis nula, por lo que son un límite superior para el error de un número potencialmente infinito de muestras que se encuentran dentro del el área especificada
es una propiedad de los datos	sì
es una propiedad de la prueba	no, se trata de la tasa de error α
dependen directamente del tamaño de la muestra	para Neyman-Pearson en el contexto de un análisis de potencia
es una expresión de los datos y una propiedad de los datos	sí
es una probabilidad (condicional)	sí (¡pero esta formulación es demasiado inespecífica!)
es la probabilidad de que los efectos observados sean aleatorios	no, los valores p son la probabilidad de los datos – dado que la hipótesis nula sea válida, pero ella no es una hipótesis real no, los valores p son la probabilidad de los datos – dado que la hipótesis nula sea válida, pero ella no es una hipótesis real
indica hasta qué punto la hipótesis nula es verdadera – dados los datos recogidos.	no, la hipótesis nula no puede demostrarse y la verdad no existe en la ciencia
indica cuán cierta es la hipótesis de investigación	no

Continuación en la página siguiente ...

indica lo cierta que es la hipótesis nula y 1 - p es la probabilidad de que la hipótesis alternativa sea cierta	no y no
es una probabilidad posterior y puede interpretarse según Bayes.	no, la interpretación bayesiana de las probabilidades de hipótesis se denota de otro modo, véase también "wishful thinking" en Gigerenzer (2004b, p.595); Sellke Bayarri & Berger (2001)
es la tasa de error α (tipo I)	no
indica la probabilidad de que los datos pueden reproducirse en x ensayos (1 - p como expresión de la fiabilidad de obtener un resultado similar).	no
indica la probabilidad de replicar exactamente ese resultado la próxima vez.	no
es una prueba de xyz...	no
dice algo sobre la importancia de los efectos encontrados	no, desde luego no por sí solo y no sin tener en cuenta los tamaños de efecto y de muestra, así como la potencia
necesita convenciones para los niveles de significación	no, sino una justificación de fondo para la selección de la probabilidad crítica de superación (para determinar la significación)
tiene más evidencia cuanto más pequeño es (rechazo de la H_0, es decir, los valores *p* pequeños son sinónimo de efectos grandes) o son una expresión del tamaño de los efectos (potencia de los efectos).	no (véase la paradoja de Jeffreys-Lindley, cap. 4.3.9), valores p muy pequeños con tamaños de muestra suficientemente grandes sólo nos dicen algo sobre el hecho de que algo que cabe esperar en determinadas condiciones - a saber, la H0 - rara vez es de esperar
indica importancia ecológica (valores p pequeños significan efectos ecológicamente importantes)	no, véase Yoccoz (1991) con un ejemplo sobre la diferencia significación estadística y significación ecológica
es una medida absoluta de la probabilidad de las hipótesis	no
sólo dice algo sobre los datos recogidos empíricamente	no, sino también sobre datos más extremos, pero no recogidos y no observados (véase la cita de Jeffreys, cap. 4.3.9)
es una propiedad de las hipótesis	no
depende del tamaño del efecto o es independiente del tamaño de la muestra	no (al aumentar el tamaño de la muestra, el tamaño del efecto no cambia, pero el valor p disminuye, ya que los errores estándar tienden a cero)
a p = 0.01 encontraremos 1 - p = 99% de todos los estudios que tienen una significación equivalente	no
depende del tamaño de la muestra en caso de que haya un efecto	sí
depende del tamaño de la muestra	sí

Tras la estadística clásica, pasamos ahora al análisis exploratorio creativo de datos según Tukey (1977).

Capítulo 5
Análisis Exploratorio de Datos (AED) según Tukey

> »Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made.«
>
> *The future of data analysis*
> John Wilder Tukey, 1962

5.1 Encontrar, buscar y revisar estructuras

Cuando se trata de explorar datos, hace falta creatividad y dejar de lado las suposiciones previas para encontrar nuevas estructuras y conexiones. Esto nos lleva al AED según John Wilder Tukey (1915-2000). Tukey (1962) describe su procedimiento como en la cita anterior y concede gran importancia a la distinción entre *análisis de datos exploratorio y confirmatorio*. Según Tukey, este último recibe demasiada atención en estadística y el primero, lógicamente, demasiado poca. En el AED no hay una teoría de pruebas definida de forma estricta ni procedimientos fijos, como el enfoque bayesiano (véase el capítulo 6) y, sobre todo, la estadística clásica (véase el capítulo 4). Se trata más bien de *encontrar, buscar y revisar y ampliar las posibilidades de interpretación*. Se admiten todas las posibilidades, como la descripción de distribuciones y sus valores característicos, todas las representaciones gráficas, los análisis de subgrupos y subconjuntos, la omisión de datos o el uso de datos de otras fuentes, omitir datos o añadir otros nuevos, transformaciones lineales y no lineales de datos, y mucho más. En Tukey (1977) se puede encontrar una gran cantidad de ejemplos elaborados.

Si consideramos lo que hay detrás del AED, se trata sobre todo de transformaciones y subagrupaciones de datos, así como de diversas visualizaciones, con el fin de ordenar los datos desde distintos puntos de vista y según criterios de interés, para estructurar o simplemente visualizar los datos. Esto es similar a la *comparación constante de casos*, un método de investigación cualitativa (véase el capítulo 9.1 o 9.5.6) y puede contarse posteriormente entre el complejo de descubrir conexiones, diferencias y estructuras, que a su vez pueden ser accesibles a futuras pruebas confirmatorias. También es concebible examinar estos datos lejos de la estadística, por ejemplo cualitativamente con el paradigma de codificación o reconstrucción (s. cap. 9 u 11) o comparativamente con el álgebra de Boole (véase el capítulo 12).

5.2 Procedimientos AED típicos en R

Algunas de las ideas de AED se remontan en el tiempo mucho antes de la época de Tukey y son ya muy antiguas. Por ejemplo, el naturalista británico, fundador de la eugenesia y escritor Francis Galton (1822-1838) ya trabajaba con la media, la mediana y la desviación estándar para obtener una medida de la desviación.

Ya utilizaba la distribución normal (véase también *Galton-Brett*, Galton, 1889, p.63, Fig.7-9), desarrolló el coeficiente de correlación – que fue elaborado matemáticamente por Pearson – e introdujo el concepto (biológico) de *regresión a la media*. Arthur Lyon Bowley (1869-1957) utilizó precursores del diagrama de tallo a principios del siglo XX y un resumen comparable a los "five numbers" para describir conjuntos de datos. Del mismo modo, la reducción de datos para destacar los puntos clave de un conjunto de datos no es una idea completamente nueva. La bibliografía cita a Andrew S.C. Ehrenberg (1926-2010) y su trabajo de 1982 sobre la reducción de datos, precedido de numerosas publicaciones desde los años cincuenta. La estadística no paramétrica ya había sido desarrollada por el estadístico de Karlsruhe Gottfried E. Noether (1915-1991) (Noether, 1967). La regresión mediana, a su vez, como método de las desviaciones absolutas más pequeñas, fue aplicada por el matemático, físico y sacerdote croata Rugjer Josip Boškovic (1711-1787) para estimar de forma robusta una regresión lineal.

R ofrece un número difícilmente encuestable de posibilidades para implementar un EDA de forma seria. Los paquetes de R específicos para AED incluyen `dlookr` (Ryu, 2019-03-16), `xda` (Karn, 2018) o `DataExplorer` (AMR, 2018). Cualquier paquete específico orientado a gráficos como `vcd, ggplot2, lattice` y muchos más pueden destacarse del conjunto de paquetes R disponibles.

A efectos analíticos, las áreas de aplicación de AED pueden subdividirse en función de sus objetivos. A continuación, hablaremos de las funciones típicas de R que cubren estos objetivos y áreas. La lista de paquetes y funciones de R que aparece a continuación es sólo una selección muy pequeña de los posibles paquetes de R. Nos concentraremos en herramientas estándar sencillas, así como en herramientas muy comunes y bien y en paquetes R muy comunes y bien soportados. Los objetivos típicos de AED son

- la *descripción de datos* (similitudes y diferencias), sus distribuciones y la exploración de estructuras ocultas y subyacentes, tanto numéricamente como gráficamente,
- el *cambio de perspectiva*, es decir, la subdivisión de conjuntos de datos según subgrupos o según aspectos de interés, variables y características, así como
- el *diagnóstico de los modelos* estimados.

Los *métodos típicos* de AED son los *estadísticos descriptivos* mediante `summary()` o `fivenum()`, especialmente los robustos como la mediana con `median()` o los cuantiles con `quantile()`. Hemos combinado los estadísticos descriptivos más comunes en la función de R `describe()` que toma vectores numéricos `c()` o matrices `matrix()` o un marco de datos `data.frame()`. Otra gran área de AED consiste en datos (no) lineales, transformaciones de datos – logaritmizar con `log()` o una transformación con las raíces cuadradas `(sqrt()`, Tukey, 1977, cap. 3 con ejemplos de datos).

Los gráficos de cualquier tipo son el foco principal de AED gráficos de dispersión para mapear relaciones (con `plot()` o extendido con `scatterplot()` del paquete R `car` o `scatterplot3d()` del paquete `scatterplot3d`). Se puede añadir cualquier información a estos gráficos (por ejemplo, líneas de regresión simples o líneas de regresión lowess con `lowess()`). Boxplots (`boxplot()`), violinplots (una mezcla de boxplot y kernel density plot, véase `vioplot()` en el paquete R `vioplot`), barplots (`barplot()`) o dotplots (`dotchart()`) o sunflowerplots (`sunflowerplot()`) proporcionan información sobre la distribución de los datos en todo el conjunto de datos o para subgrupos interesantes.

La subdivisión en subgrupos (por ejemplo, grupos de edad, sexo, según categorías cualitativas, etc.) puede hacerse manualmente o con paquetes de R adaptados, como `ggplot2` o `lattice`, que están especializados en ello y han adaptado a los subgrupos muchas de las variantes de trazado que se enumeran aquí. En el sitio web *The R Graph Gallery* se puede encontrar una inspiradora visión general de esto y de las capacidades gráficas de R.

Los datos multivariantes o los datos de frecuencia (por ejemplo, muchas categorías cualitativas) pueden trazarse utilizando heatmaps (`heatmap()` o `heatmap3()` del paquete R `heatmap3`, `heatmap.plus()` de `heatmap.plus` o interactivamente con `heatmaply()` de `heatmaply`),mosaicplots (mosaicplot() o `mosaic()` del paquete R `vcd`), scatterplots múltiples (`splom()` del paquete R `lattice`, donde `SPLOM` = matriz scatterplot), correlogramas (`corrgram()` del paquete R `corrgram`), que son matrices de dispersión que se centran en las relaciones correlativas y similares. Las matrices de dispersión son adecuadas para obtener histogramas y estimaciones de densidad en las diagonales y diagramas de dispersión (con líneas de regresión)

y correlaciones (numéricas, como elipses, codificadas por colores) en las diagonales secundarias. Los paquetes de R `vcd` y `vcdExtra` ofrecen más posibilidades para visualizar datos categóricos (por ejemplo, `assoc(), cd_plot(), cotabplot(), doubledecker() fourfold(), spine()`).

A nivel de diagnóstico, los rootogramas (`rootogram()` en el paquete `vcd` de R) muestran el ajuste de de los datos de frecuencia a distribuciones (por ejemplo, la distribución de Poisson, véase Tukey, 1949 o Kleiber & Zeileis, 2016). En este contexto, los dendrogramas, es decir, las visualizaciones de los análisis de conglomerados jerárquicos (`hclust()`), pueden perfeccionar la imagen y los aspectos y correlaciones interesantes y pueden ayudar a descubrir aspectos y correlaciones interesantes. Lo mismo se aplica a la salida gráfica del escalado multidimensional (`cmdscale()`), el análisis de componentes principales (`princomp()`) o un gráfico de dispersión de los resultados de un análisis lineal discriminante `lda()` en el paquete `MASS` de R). Histogramas (`hist()`) y (kernel) estimaciones de densidad (`density()`), por ejemplo, complementan los gráficos de caja mostrando la distribución de los datos (cuerpos de datos, extremos, valores atípicos, etc.) de forma muy precisa. Sombreado coloreado, por ejemplo con `smoothScatter()` para gráficos de dispersión o dentro de otros gráficos (por ejemplo, histogramas `hist()`) ayuda con la orientación – por ejemplo, para subgrupos, para determinados rangos de valores, para pruebas previas o posteriores, series temporales, etc. Los gráficos de interacción `interaction.plot()` son adecuados para conocer las dependencias de ciertas variables respecto a otras, por ejemplo, cuando se trata de los valores de las variables dependientes en relación con los niveles de los factores experimentales. Representar gráficamente los datos con respecto a determinadas distribuciones (por ejemplo distribución normal con `qqnorm(), qqline()` y `qqplot()` o extensiones en el paquete R `car`) ofrece una visión de prueba en el contexto de exploraciones de datos o análisis residuales y contrasta la expectativa con la realidad. Los datos pueden representarse gráficamente con respecto a la distribución normal o a otra distribución. Dadas las cualidades gráficas superiores de programas de análisis estadístico como R, o dadas las capacidades comparables de lenguajes de programación como Python, los métodos anteriores favorecidos por Tukey como los gráficos de tipo "stem-and-leaf" (tronco y hoja, `stem()`, que están en el rango ASCII) son cada vez menos comunes, pero siguen siendo útiles. Los diagnósticos de modelos lineales simples – como con el paquete R `car` – se basan más ampliamente en gráficos de los datos estimados por el modelo y predichos por él frente a las expectativas teóricas u otros conjuntos de datos en el caso de la predicción. En el análisis de modelos lineales, los gráficos de residuos estandarizados (=si se conoce la verdadera desviación estándar de los residuos) o studentised (=si se desconoce la verdadera desviación típica de los residuos, se toma la desviación estándar de la muestra empírica de los residuos) frente a los valores ajustados (=valores estimados por el modelo) son valiosos y guían la búsqueda de sesgos, valores atípicos y variables de talla única. La salida gráfica de un modelo lineal generado con `lm()` ya muestra la mayoría de estos diagnósticos por defecto, como demuestra de forma ejemplar la página de ayuda de `lm()` con la aplicación del conjunto de datos de Annette Dobson (1990, s. cap. 4.4.10 o 4.4.11) sobre el peso de diferentes especies de plantas (`ptII_quan_EDA_intro_-overviewrobust.r`):

```
example(lm)
```

Las cantidades de influencia distorsionantes o las irritaciones en el conjunto de datos – por cualquier motivo – pueden dilucidarse con relativa facilidad mediante gráficos en combinación con determinados coeficientes (por ejemplo, la distancia de Cook, por ejemplo en combinación con un gráfico de influencia, `influencePlot()` del paquete R `car`). Lo mismo se aplica a la búsqueda de colinealidades y factores de inflación de varianza (Fox, 2002, vif() del paquete R `car`) o varianzas de error no constantes (`SpreadLevelPlot()` del paquete R `car`). Fox (2002) o Fox y Weisberg (2019) ofrecen una muy buena visión general de los diagnósticos de los modelos lineales, y Kabaco (2017) la resume muy brevemente con un código R sencillo.) Para modelos lineales jerárquicos o generalizados más complejos, existen los paquetes de R `nlme, lme4, HLMdiag` y `sjPlot`. De forma equivalente, se pueden encontrar gráficos de diagnóstico para probabilidades posteriores y estimaciones de parámetros en el curso de análisis bayesianos (por ejemplo, `boa, coda, MCMCvis`, véase también la descripción general de los paquetes `Bayes` en R en CRAN, 2019a).

Este extracto se puede complementar con muchos otros métodos, por ejemplo, de forma interactiva mediante el paquete R rggobi ().

En resumen, la idea detrás de AED es utilizar transformaciones de datos, subagrupaciones y visualizaciones para ordenar y estructurar los datos desde diferentes perspectivas y según criterios de interés con el fin de hacer accesibles las estructuras contenidas en ellos.Debemos tener cuidado de no seguir solo las normas y estándares establecidos cuando se trata de ciencia y, por lo tanto, siempre de lidiar con la incertidumbre real. Pero, como cualquier otro análisis, el AED no pretende hacer afirmaciones arbitrarias, sino coherentes y estables – análisis robustos.

5.3 Análisis robusto de datos como parte del AED

Muchos de los métodos introducidos por Tukey (por ejemplo, el boxplot, Tukey, 1977) son muy prácticos, perseguen el objetivo de robustez e implican no sólo un resumen puntual de una distribución, sino unas cuantas medidas robustas de toda la distribución. Aunque el tiene sus trampas y reacciona de forma sensible a los valores atípicos extremos, la mediana no presenta este problema y, a pesar de utilizar menos información que la media, estima una distribución sorprendentemente bien. Mostramos un ejemplo con datos aleatorios de una distribución normal, que lo demuestra bien (`ptII_quan_EDA_intro_overviewrobust.r`).

```
> # median vs. mean
> mu <- 6
> sigma <- 2.34
> N <- 1e4
> seed <- 54321
> reps <- 1e3
> res.2 <- mwmed.sim(N=N, mu=mu, sigma=sigma, reps=reps,
                   + abso=FALSE, pr=TRUE, usesameseed=FALSE)
                     Min.        1st Qu.      Median        Mean
mw               5.57149865  5.973551428  6.001799e+00  6.0006427705
med              5.67669631  5.966303515  5.999879e+00  6.0016199505
ratio.medmw     -0.09591891 -0.003116319 -8.352224e-07 -0.0001687402
dev.mw          -0.18312024 -0.011302808  7.688924e-04  0.0002746883
dev.med         -0.13816397 -0.014400207 -5.169374e-05  0.0006922865
                     3rd Qu.     Max.        SD                 VAR
mw                   6.026611027 6.36536848 0.054057279 2.922189e-03
med                  6.033032374 6.94092252 0.068398820 4.678399e-03
ratio.medmw          0.003217613 0.03616831 0.007701615 5.931487e-05
dev.mw               0.011372234 0.15614038 0.023101401 5.336747e-04
dev.med              0.014116399 0.40210364 0.029230265 8.544084e-04
> mwmed.sim.plot(res.2,mu=mu)
```

La función anterior `mwmed.sim()` genera datos aleatorios distribuidos normalmente para tamaños de muestra crecientes con parámetros constantes (media, desviación típica, valor inicial idéntico para el generador aleatorio). La relación entre la mediana y la media se calcula para cada muestra y se emite como una relación de desviación de uno, además de las desviaciones de la media muestral x y de la mediana muestral del valor de la población μ, normalizadas a la desviación estándar σ de la población. La función R muestra estadísticas descriptivas y les devuelve con los datos brutos. Las llamadas posteriores trazan estas desviaciones con `mwmed.sim.plot()` (véase la Fig. 5.1). La línea horizontal verde en los gráficos marca una desviación de cero o la línea de regresión entre las medianas y las medias empíricas. Como puede verse, la diferencia entre la mediana y la media disminuye a medida que aumenta el tamaño de la muestra, a pesar de ocasionales mayores fluctuaciones de la muestra. La llamada con `set.seed()` permite una reproducción utilizando el valor inicial del propio generador aleatorio de R. El paso de `abso=FALSE` requiere que no se devuelvan valores absolutos, sino desviaciones mayores y menores que cero. Los lectores interesados pueden repetirlo todo con una muestra aún mayor, por ejemplo n = 1e5 o mayor, entonces la progresión hacia

condiciones infinitas se hace mucho más clara. En `usesameseed=FALSE` asegura que para cada simulación se elige un valor inicial diferente para el generador aleatorio. No obstante, es interesante observar simplemente la progresión con el mismo valor inicial.

Por las razones anteriores, la regresión robusta ("línea mediana-mediana") se derivó de la mediana (Walters, Morrell & Auer, 2006). Tukey utilizó medidas de dispersión y varianza para hacer estimaciones de la distribución de los datos, como el resumen del "five numbers" formado por el mínimo, el máximo, la mediana y los cuartiles inferior o superior (1er y 3er cuartil). En R, estos valores son accesibles mediante `fivenum()`. Estos parámetros robustos se denotan para todas las distribuciones según Tukey, que no se aplica a la media ni a la desviación estándar. Así, por ejemplo, estas dos características no pueden calcularse para la distribución de Cauchy o la de Lorentz. Además, estas cinco características de distribución resultan ser mucho más robustas frente a las distorsiones que pueden surgir de distribuciones sesgadas o de colas pesadas en los extremos. Un ejemplo es la distribución *t* con pocos grados de libertad en comparación con la distribución normal. El uso de la distribución *t* en lugar de la distribución normal – por ejemplo, con muestras más pequeñas – representa un intento de analizar de forma más robusta.

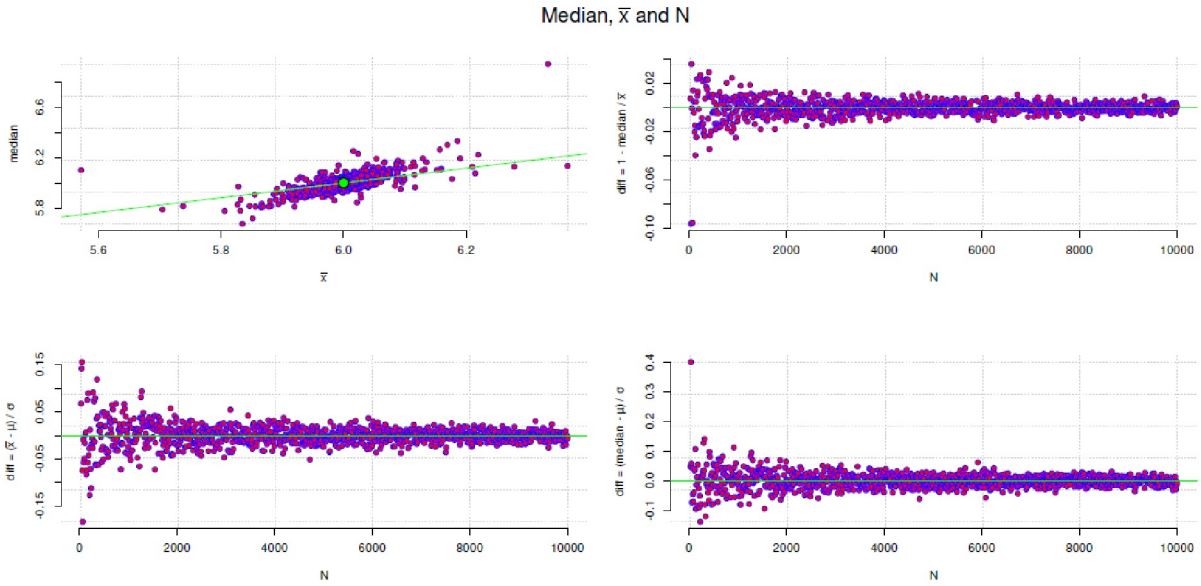


Figura 5.1. *Relación entre la mediana y la media (tamaño de la muestra aumentando)*

Por un lado, el uso de estadísticas robustas supone, en sentido estricto, una cierta pérdida de información, ya que se utiliza menos información en el cálculo de los valores característicos. Por otra parte es mucho menos probable que los extremos distorsionen las tendencias centrales (por ejemplo, sobre la media), por lo que los métodos robustos no sólo son legítimos, sino incluso necesarios para explorar los datos, estimar los parámetros y amplar las medidas comunes por defecto. Tampoco se trata simplemente de incluir o excluir procedimientos, sino de aplicar distintos métodos de análisis de datos para que los datos y el proceso de generación subyacente pueden comprenderse mejor. Posteriormente, se selecciona el modelo adecuado sobre una base de información más amplia con el fin de extraer conclusiones. De esta manera la información disponible utiliza de forma adecuada y completa. Los procedimientos robustos no sólo se utilizan descriptivamente sino también para construir modelos y pruborlos. Por tanto, *el análisis robusto de datos* consiste en identificar los valores atípicos y otros parámetros que distorsionan los datos, su influencia en el conjunto de datos y minimizarlos si es posible o, si es necesario, excluirlos al menos a modo de prueba.

Por ejemplo, se pueden asignar ponderaciones a datos distintos para que los valores atípicos reciban menos influencia (por ejemplo, `rlm()` en `MASS`) o se corten partes de los datos en los extremos. Se considera que los métodos robustos son (más) independientes de los requisitos habituales para el uso de modelos `OSL`

(OSL = orinary least squares / mínimos cuadrados ordinarios, es decir, mínimos cuadrados simples), por ejemplo en en cuanto a la necesaria homogeneidad de la varianza. Sin embargo, a veces hay formas sencillas de trabajar de forma más robusta con modelos OSL, como muestra Tofallis (2008), que trabaja con errores porcentuales en lugar de errores absolutos, y así datos atípicos mantienen su posición externa, pero en condiciones mucho menos extremas. Este procedimiento se aplica después a la regresión OSL. Según el autor, esto conduce además a una actitud robusta frente a las violaciones de la homocedasticidad. Suponiendo la normal de estos errores relativos, se obtiene una estimación de Máxima Likelihood.

Por otro lado, ¿se puede justificar de forma sustancial que los valores atípicos provienen de un proceso de generación de datos diferente al del grueso de los datos y, por eso, se pueden modelar por separado los valores atípicos o incluso eliminarlos? Esto permite un cierto grado de solidez en caso de que se conserven los métodos de análisis de datos utilizados, pero requiere una justificación sustantiva sólida. Los datos no pueden excluirse simplemente porque no gusten (o ya no gusten) sobre la base del modelo elegido. Las consecuencias de utilizar métodos no robustos son un aumento de la tasa de error de tipo II, una disminución de la tasa de error de tipo I y, en consecuencia, una sobreestimación de la anchura de los intervalos de confianza en la estimación de parámetros. La robustez a corto plazo en el análisis de datos significa realizar pruebas mucho más conservadoras y pasar por alto o no tomar suficientemente en serio los posibles efectos.

La idea que subyace a la estimación robusta es que – para una estimación robusta – los datos entran ponderados en el modelo: a los datos extremos se les da menos peso o, en casos extremos, ningún peso. Los datos con residuos no iguales a CERO reciben un peso - para desviaciones pequeñas igual a UNO o, de haber una desviación, un peso inferior que UNO. Esto varía en función del algoritmo utilizado para determinar la ponderación, que está directamente relacionado con el procedimiento de estimación global del parámetro o modelo de interés. Procedimientos conocidos para determinar las ponderaciones son Huber *k* (Huber, 1981), Hampel (Hampel, Ronchetti, Rousseeuw & Stahel, 1986), Andrews Wave (Andrews, 1991) o Tukey's biweight (Mosteller & Tukey, 1977).

Una comparación sencilla de la estimación OSL y robusta en R puede realizarse utilizando `lm()` y `rlm()` del paquete MASS de R. Otros paquetes de R, como `robust, quantreg, robustlmm` o `robustbase` están disponibles y ofrecen otras funciones de R muy amplias sobre el tema de la estimación robusta. Para `rlm()` hay varios de los algoritmos de ponderación antes mencionados y se pueden selccionarlos mediante el parámetro `psi` (`psi.huber, psi.hampel` y `psi.bisquare`). Algo comparable es posible en el paquete R `robustbase`. Echemos un vistazo a estos algoritmos de ponderación de forma gráfica:

```
# functions to plot psi functions
# https://en.wikipedia.org/wiki/Robust_statistics
# https://de.wikipedia.org/wiki/Datei:Mest_weightfunc.jpg
andrewswave <- function(sek, a=1.34*pi)
{
sek.abs <- abs(sek)
sek.mal.pi <- pi*sek
a / (sek.mal.pi) * sin(sek.mal.pi/a) * ( sek.abs <= a )
}
# graphical plot of different weighting functions
sek <- seq(-10,10, length.out=1e3)
digs <- 3
par(oma=c(2,2,2,2), mar=c(2,2,4,2), "cex.axis"=1,
           bty="l", mfrow=c(2,2))
# Huber k
plot(sek, psi.huber(sek), type="l", col="darkred", bty="n",
           pre.plot=grid(), main="Huber k", ylab="weight")
# Hampel
plot(sek, psi.hampel(sek), type="l", col="darkred", bty="n",
pre.plot=grid(), main="Hampel", ylab="weight")
# Andrews Wave
a <- 1.34*pi
plot(sek, andrewswave(sek, a), type="l", col="darkred", bty="n",
           pre.plot=grid(), ylab="weight",
           main=expression(paste("Andrews Wave (a = 1.34 * ",pi,")",sep="")))
# Tukey's biweight
plot(sek, psi.bisquare(sek, a), type="l", col="darkred",
```

```
bty="n", pre.plot=grid(), main="Tukey's biweight", ylab="weight")
mtext(expression(paste("Various weighting functions",sep="")),
          outer=TRUE, line=-1, cex=1.5, side=3)
```

Como puede observarse (véase la Fig. 5.2), la influencia de los valores extremos disminuye constantemente con Huber k, mientras que con Hampel, AndrewsWave y el biweight de Tukey los valores muy extremos reciben un peso de CERO. Si ahora se imagina una distribución de datos y sus residuos y se multiplican los datos de la imaginación por la distribución de ponderación visualizada, queda claro con un poco de creatividad que las funciones de ponderación elegidas pueden minimizar o incluso eliminar la influencia de los valores extremos. Sin embargo, al utilizar funciones de ponderación, hay que tener cuidado de que los datos ya no se ponderen por igual en el modelo, sino que se produzca una selección y, por tanto, una distorsión – deseada, pero aún así una distorsión de los datos originales. Ahora, por supuesto, podemos preguntarnos: "¿Cómo son los datos realmente disponibles?", "¿Qué relevancia tienen esos datos para mi pregunta?", etc.? Todas estas son preguntas que pueden hacerse inmediatamente cada vez que construimos una escala, nos preguntamos por la logaritmización de los datos, etc., al igual como con el tema de la robustez de las estimaciones.

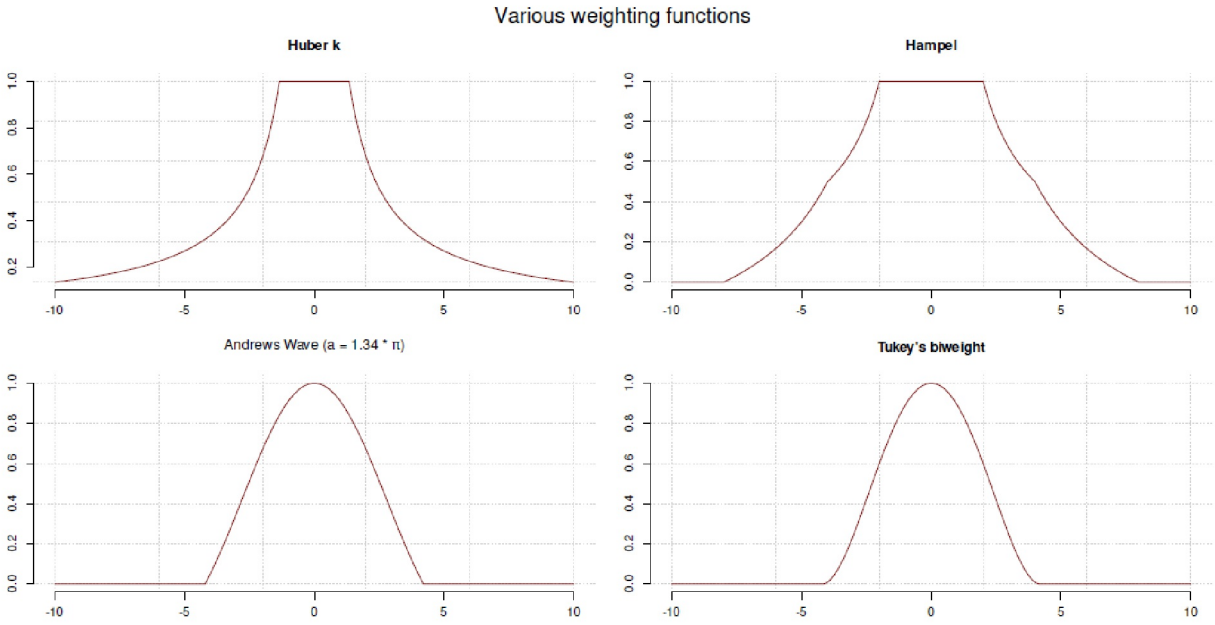


Figura 5.2. *Funciones de ponderación para análisis robustos*

Toda la estimación del modelo se determina siempre de forma iterativa y requiere un valor inicial significativo, por ejemplo la mediana. Por eso hablamos de una estimación IRLS (= iteratively reweighted least squares / mínimos cuadrados iterativamente reponderados). Pertenecen a la clase de los estimadores M (M = maximización), que son estimadores del tipo de maximum likelihood /máxima verosimilitud (ML) y generalizan ML (Huber 1981), pero son más robustos que éstos. No es difícil comprender que existen diferentes enfoques para estos problemas de estimación y que, como de costumbre, no hay una solución unívoca y válida para siempre. Mirando más allá, están los estimadores L, R, S y MM (Jurecková, 1984; Rousseeuw & Yohai, 1984; Susanti, Pratiwi, Sulistijowati & Liana, 2014), que no discutimos aquí. Como alternativa a estos métodos de estimación, se puede utilizar, por ejemplo, el algoritmo RANSAC (= random sample consensus / consenso de muestras aleatorias, Fischler & Bolles, 1980), que también funciona de forma iterativa y no intenta equilibrar todos los datos en paralelo, sino sólo tantos (aleatoriamente) como sea necesario para calcular los parámetros del modelo. El algoritmo tiene su origen en trabajos sobre reconocimiento de imágenes. El supuesto de robustez se comprueba contrastando los residuos de todos los datos con el modelo estimado. Además, se fija un umbral crítico para saber cuánto puede desviarse un dato

del modelo para que no se considere un error en el sentido de un valor atípico perturbador, sino un representante legítimo del modelo estimado. Esta secuencia de pasos se lleva a cabo por separado para cada iteración, y a continuación los resultados se integran por consenso y se estiman de nuevo utilizando métodos tradicionales.

Las desventajas son, por ejemplo, que el algoritmo no siempre encuentra la solución óptima si hay muy pocas iteraciones, que el número de iteraciones necesarias no está limitado al alza y que puede fallar incluso con datos moderadamente contaminados si hay muy pocos representantes legítimos ("inliers" $< 50\%$ frente a "outliers") en los datos. Como resultado, ha habido extensiones y desarrollos posteriores de RANSAC en el contexto del reconocimiento de imágenes y patrones para mejorar los problemas respectivos.

La robustez de los estimadores se indica generalmente mediante el punto de ruptura. Éste indica la proporción de datos que se necesita para que un estimador se colapse, es decir, para distorsionar arbitrariamente el resultado de la estimación. Para la media aritmética, el punto de ruptura es aproximadamente $n = 1$. Esto significa que un solo valor atípico grande ya es suficiente para distorsionar arbitrariamente el valor medio. Para las estimaciones robustas, se puede alcanzar un punto de ruptura de 0.5, que representa el máximo. Esto se debe a que si más del 50% de los datos están contaminados, ya no es posible determinar claramente cuál es la distribución real subyacente a los datos. La mediana, por ejemplo, tiene un punto de ruptura de 0.5. Si se recortan los datos, es decir, si se elimina el x% de los datos, el punto de ruptura es x%.

Para ilustrar lo que hemos descrito, elegimos una regresión robusta y tomamos prestado el conjunto de datos `crimedata` (Agresti & Finlay, 1997) para el análisis de la delincuencia. Nos limitamos a dos variables como predictores, a saber, la pobreza y la monoparentalidad. La variable dependiente es la delincuencia violenta por cada 100.000 habitantes (`ptII_quan_EDA_intro_overviewrobust.r`). Calculamos una estimación `OLS` y una estimación robusta. Por defecto, `rlm()` utiliza la ψ-función `psi=psi.huber` para determinar las ponderaciones:

```
# lm vs. rlm
# Crime data
# explanations
# https://stats.idre.ucla.edu/r/dae/robust-regression/
# read from the net
# URL <- c("https://stats.idre.ucla.edu/stat/data/crime.dta")
# crimedata <- read.dta(URL)
# write.table(crimedata, file="crimedata.tab", sep="\t",
            row.names=FALSE, col.names=TRUE)
crimedata <- read.csv(file="crimedata.tab", sep="\t", header=TRUE)
head(crimedata)
tail(crimedata)
crimedata.lm <- lm(crime ~ poverty + single, data = crimedata)
crimedata.rlm <- rlm(crime ~ poverty + single,
            data = crimedata, psi = psi.bisquare)
summary(crimedata.lm)
summary(crimedata.rlm)
```

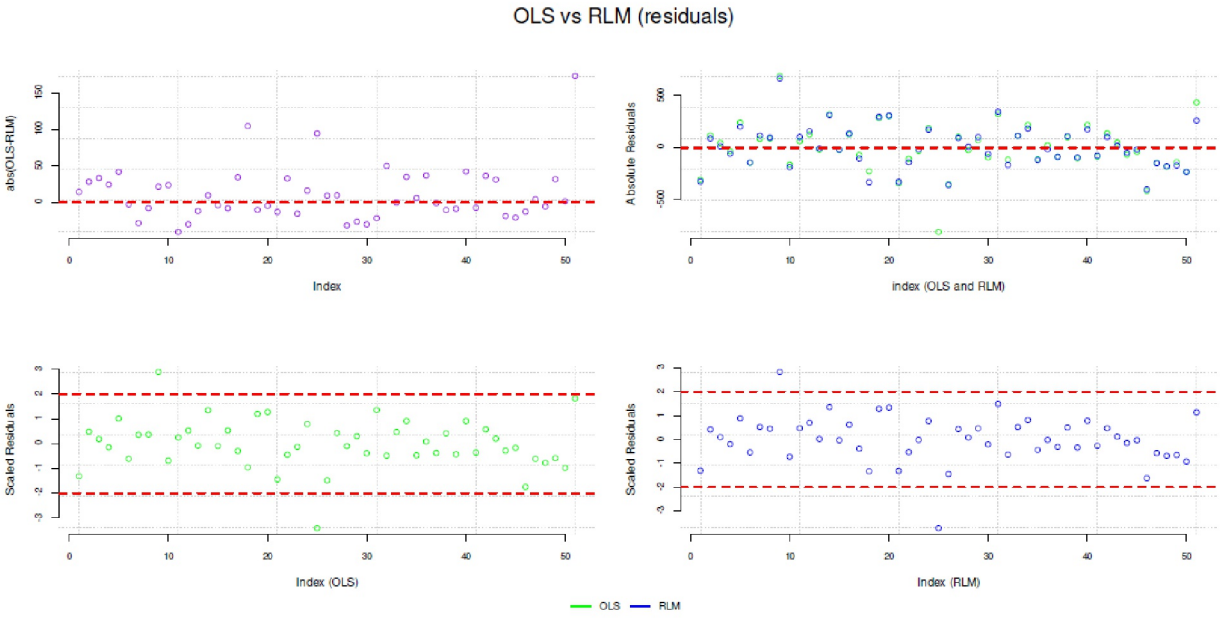


Figura 5.3. *Datos de criminalidad (gráficos de residuos)*

A continuación, formamos un índice de eficiencia relativa relacionando los dos errores estándar residuales entre sí. Los residuos pueden representarse gráficamente (véase la Fig. 5.3), de modo que podemos ver la diferencia entre los OLS y la estimación robusta:

```
> # relative efficiency
> summary(crimedata.lm)$sigma / summary(crimedata.rlm)$sigma
[1] 1.203957
> # residual plot
> resid.plot(lmfit=crimedata.lm, rlmfit=crimedata.rlm)
```

Del mismo modo podemos observar cómo difieren los coeficientes entre los dos modelos:

```
> # ratios of coefficients lm vs. rlm
> coef(crimedata.lm)/coef(crimedata.rlm)
(Intercept) poverty    single
0.8911344   0.5805960 0.9456737
> coef(crimedata.lm)/( coef(crimedata.lm) + coef(crimedata.rlm) )
(Intercept) poverty    single
0.4712168   0.3673272 0.4860392
> # the higher the residual, the lower the weight
```

Se aplica a los residuos que cuanto mayores son, menor es su ponderación en el modelo robusto (salida abreviada):

```
# the higher the residual, the lower the weight
data.frame(crimedata.rlm$residuals, crimedata.rlm$w)
crimedata.rlm.residuals crimedata.rlm.w
1           -325.851891   0.777662383
2             88.723307   0.982552633
3             12.072733   0.999674926
...
50          -234.195152   0.881660897
51           260.648854   0.854441716
```

También podemos verlo gráficamente (véase la Fig. 5.4):

```
# how are weights distributed? vs. residuals
with(crimedata.rlm, plot(residuals, w, bty="n", pre.plot=grid(),
col="darkred", ylab="Weights", xlab="Residuals"))
abline(h=1, v=0, col="blue", lty=2)
mtext("Distribution of weights vs. residuals",
outer=TRUE, line=-3, cex=1.5)
```

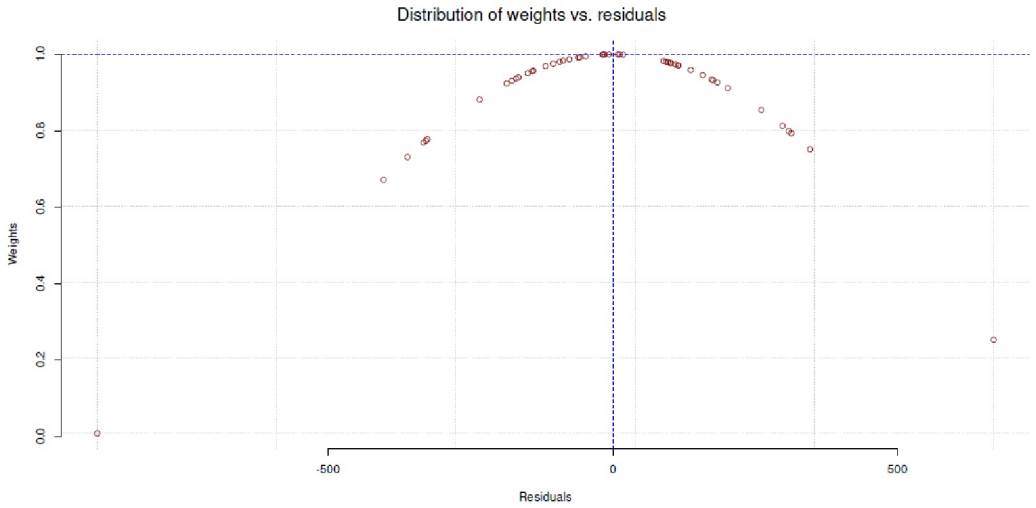


Figura 5.4. *Datos de criminalidad (ponderaciones vs. residuos)*

La diferencia entre OLS y robusta no siempre parece clara, como demuestra el trabajo con el conjunto de datos `bupa`. Los datos contienen marcadores relacionados con las enfermedades hepáticas y el consumo excesivo de alcohol (Breiman, 2001), aquí en relación con los hombres estudiados. Examinamos dos variables, a saber, la alamina aminotransferasa (`sgpt`) y la gamma-glutamil transpeptidasa (`gammagt`). Para el diagrama de dispersión, se logaritman ambas variables y se estima un modelo lineal de la forma `log(sgpt) ~ log(gammagt)` (véase la Fig. 5.5).

```
# BUPA data liver disorders
# read from the net
# URL <- "ftp://ftp.ics.uci.edu/pub/machine-learning-databases/
liver-disorders/bupa.data"
# bupa <- fread(URL)
# colnames(bupa) <- c("mcv","alkphos","sgpt","sgot", "gammagt","drinks","selector")
# bupa$gammagt.log <- log(bupa$gammagt)
# bupa$sgpt.log <- log(bupa$sgpt)
# write.table(bupa, file="bupa.tab", sep="\t", row.names=FALSE, col.names=TRUE)
bupa <- read.csv(file="bupa.tab", sep="\t", header=TRUE)
head(bupa)
tail(bupa)
# OLS
bupa.lm <- lm( sgpt.log ~ gammagt.log, data=bupa)
# RLM
bupa.rlm <- rlm( sgpt.log ~ gammagt.log, data=bupa)
# output
summary(bupa.lm)
summary(bupa.rlm)
# scatterplot
with(bupa, plot(gammagt.log, sgpt.log, col="purple", bg="grey", xlab="log(gammagt)",
           ylab="log(sgpt)", pch=21, lwd=2, pre.plot=grid(), bty="n"))
abline(bupa.lm, col="green", lwd=2)
abline(bupa.rlm, col="blue", lwd=2, lty=2)
```

```
legend("topleft", legend=c("OLS","RLM"), col=c("green","blue"),
           bty="n", lty=c(1,2), lwd=c(2,2))
mtext("log(sgpt) vs. log(gammagt)", outer=TRUE, line=-3, cex=1.5)
```

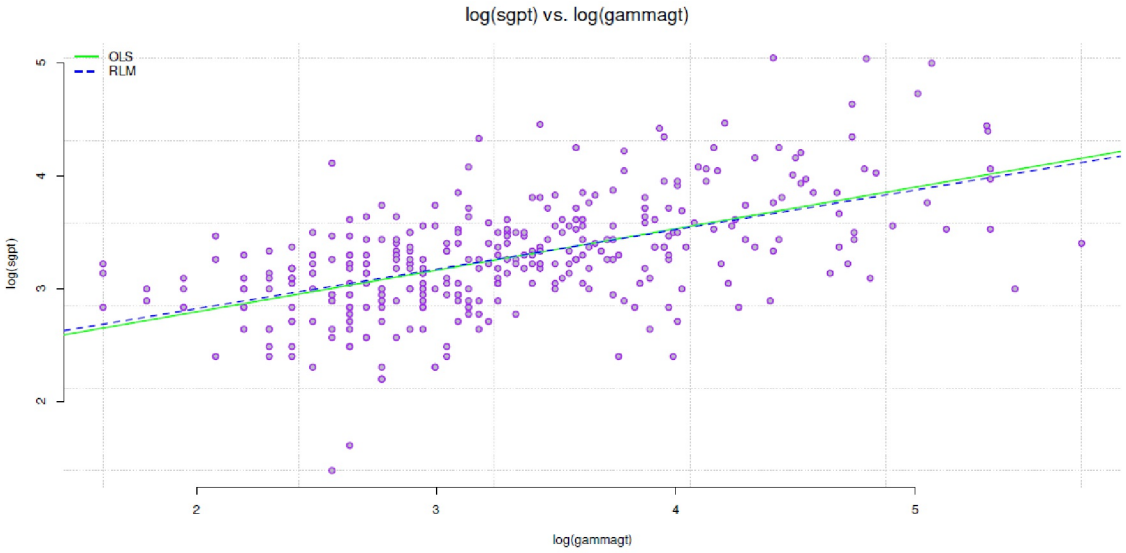


Figura 5.4. *Enfermedades hepáticas y consumo de alcohol*

Mientras que las líneas de regresión apenas difieren visualmente, otra mirada a la relación de los errores estándar residuales muestra que la relación de los errores estándar residuales, sin embargo (gráfico no impreso) un valor de 1:183. Esto significa que el OLS es 1:2 veces menos excéntrico que la estimación robusta.

```
> # relative efficiency
> # the smaller the better for the first one
> # in case of one = same efficiency
> summary(bupa.lm)$sigma / summary(bupa.rlm)$sigma
[1] 1.183247
> # residual plot to identify outliers
> resid.plot(lmfit=bupa.lm, rlmfit=bupa.rlm)
```

Como regla general, si hay una gran variación en los parámetros entre OLS y la estimación robusta, debe preferirse la estimación robusta. Lo mismo puede hacerse con los datos del conjunto de datos de Chicago del libro de Faraway (2002-07). Se puede utilizar el paquete R `faraway`. El conjunto de datos se describe detalladamente y se examina con varios métodos análisis en Faraway (ibíd., capítulos 12 y 13). Contiene datos sobre la disponibilidad de seguros en distintos barrios en función de variables como incendios (fire), robos (robo), composición de la población (raza), edad (edad) e ingresos (ingresos). La variable dependiente es `involact`, que describe las pólizas y renovaciones del plan FAIR. Se trata de una medida de la libre disponibilidad de pólizas de seguro, ya que estas pólizas sólo suelen contratarse después de haber sido rechazadas en el mercado libre. Rara vez representan una preferencia de los asegurados por este tipo de pólizas. Aquí una comparación entre OLS y un modelo robusto de la forma `involact ~ raza + incendio + robo + edad + log(ingresos)` conduce a una eficacia relativa de OLS frente a robusto de 1:343.

```
> # Chicago data
> data(chicago)
> head(chicago)
race  fire theft age volact involact income
60626 10.0 6.2   29  60.4   5.3 0.0  11744
60640 22.2 9.5   44  76.5   3.1 0.1  9323
...
> tail(chicago)
```

```
race  fire theft age volact involact income
60655 1.0   4.8  19  15.2   13.0 0.0 13323
60643 42.5 10.4  25  40.8   10.2 0.5 12960
...
>
> chic.lm <- lm(involact ~ race + fire + theft + age +
+ log(income), data=chicago)
> chic.rlm <- rlm(involact ~ race + fire + theft + age +
+ log(income), data=chicago)
> summary(chic.lm)
Call:
lm(formula = involact ~ race + fire + theft + age + log(income),
data = chicago)
Residuals:
Min        1Q     Median  3Q      Max
-0.85393 -0.16922 -0.03088 0.17890 0.81228
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.573976 3.857292   -0.927  0.359582
race           0.009502 0.002490    3.817  0.000449 ***
fire           0.039856 0.008766    4.547  4.76e-05 ***
theft         -0.010295 0.002818   -3.653  0.000728 ***
age            0.008336 0.002744    3.038  0.004134 **
log(income)    0.345762 0.400123    0.864  0.392540
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.3345 on 41 degrees of freedom
Multiple R-squared: 0.7517, Adjusted R-squared: 0.7214
F-statistic: 24.83 on 5 and 41 DF, p-value: 2.009e-11
> summary(chic.rlm)
Call: rlm(formula = involact ~ race + fire + theft + age + log(income),
data = chicago)
Residuals:
Min       1Q       Median  3Q      Max
-0.99015 -0.16183 -0.01694 0.17569 0.89655
Coefficients:
            Value   Std. Error t value
(Intercept) -2.9257 3.3974     -0.8612
race         0.0079 0.0022      3.5828
fire         0.0459 0.0077      5.9404
theft       -0.0097 0.0025     -3.9124
age          0.0064 0.0024      2.6509
log(income)  0.2829 0.3524      0.8029
Residual standard error: 0.2492 on 41 degrees of freedom
```

Un examen de los coeficientes también muestra una clara diferencia entre la estimación OLS y la estimación robusta:

```
> # ratio lm vs. rlm
> coef(chic.lm) / coef(chic.rlm)
(Intercept) race      fire      theft     age       log(income)
1.2215634   1.2095301 0.8689697 1.0601679 1.3010474 1.2220124
> # relative efficiency lm vs rlm
> summary(chic.lm)$sigma / summary(chic.rlm)$sigma
[1] 1.342635
```

En CRAN (2018) se puede encontrar una descripción exhaustiva de los procedimientos robustos en R.

5.4 Diferenciar el AED de los enfoques confirmatorios

AED pretende explorar nuevas perspectivas y comparar datos en diferentes contextos en función de sus características. Esto implica especificar medidas robustas de tendencia central y de dispersión de tal manera que se obtenga una buena visión de los datos en su conjunto. Hay que identificar tendencias, patrones ocultos, valores atípicos, subgrupos interesantes, etc. y situarlos en un contexto significativo entre sí. La visualización y evaluación de datos mediante métodos gráficos (por ejemplo, Garry, Simpson, Vehtari, Betancourt & Gelman, 2019) permite un acercamiento intuitivo a los datos, hecho que ahora se utiliza ampliamente para la comprobación de modelos. Para HLMs/MLMs, esto es descrito por Loy, Hofmann y Cook (2016). Los paquetes R asociados son `HLMdiag` y `DHARMa`. En particular, la representación gráfica de subgrupos uno al lado del otro – por ejemplo, en cuadrículas (*grids*) – es un excelente método de visualización para estudiar y aprender del comportamiento específico de los datos en diferentes contextos. Lo mismo se aplica a la representación gráfica del modelo estimado frente a otros datos, como una comparación de datos simulados frente a datos empíricos.

Hoy en día, se suele denominarse el AED con grandes conjuntos de datos "*minería de datos*" (*data-mining*) y se aplica el enfoque para generar o derivar conocimiento (semi)automáticamente a partir de grandes cantidades de datos. Según la entrada correspondiente del diccionario, el conocimiento incluye la búsqueda de reglas y leyes, así como de correlaciones ocultas. Esto incluye el uso de métodos estadísticos para grandes conjuntos de datos, por ejemplo en el contexto de los análisis de bases de datos, como hacen muchas empresas. Es de suponer que esto se verá cada vez más apoyado por algoritmos de IA y posiblemente sustituida en el futuro. A veces da la impresión de que los resultados de los procesos de minería de datos son *hechos probados*, como si hubieran sido sometidos a rigurosas pruebas y exámenes críticos. Esto *no es en absoluto así*, aunque puedan descubrirse correlaciones interesantes. Más bien, los resultados son de naturaleza *puramente* exploratoria e implican, desde luego, *ninguna* planificación previa *ni* prueba meticulosa baja condiciones experimentales reproducibles. Especialmente con grandes conjuntos de datos, esto podría asemejarse incluso a un enfoque de escopeta que "ya encuentra algo" que podría parecer significativo, especialmente si se añaden pruebas de significación convencionales. El propio Tukey advirtió contra la realización de pruebas confirmatorias de las estructuras exploratorias encontradas en el mismo conjunto de datos. Esto sólo conduce a distorsiones sistemáticas y profecías autocumplidas, si los patrones encontrados mediante AED se prueban a su vez sobre sí mismos. AED conduce necesariamente a la demanda de *replicación* – ¿cómo podría ser de otra manera? Y esto también se aplica a los procesos de minería de datos. Debe observarse la separación entre exploración y confirmación, por ejemplo, subdividiendo aleatoriamente el conjunto de datos para desarrollar el modelo en una parte y probarlo en otra. Con grandes conjuntos de datos, tal subdivisión de este tipo parece posible sin problemas.

El AED también debe distinguirse del *análisis inicial* de datos en el contexto de las pruebas estadísticas. Este último sólo tiene por objeto comprobaciones estrechamente definidas y *no* pretende descubrir estructuras. Por supuesto, las suposiciones previas sobre los supuestos distribucionales y los requisitos previos de las pruebas, los datos que faltan, las transformaciones de datos necesarias, etc. se tienen en cuenta en una fase temprana. Se utilizan algunos de los mismos procedimientos que en el AED, pero la intención y la flexibilidad son claramente diferentes. Un primer análisis de los datos en el contexto de un diseño de pruebas estrictamente deductivo va seguido de pruebas confirmatorias basadas en hipótesis. La comprobación preliminar de los datos tiene por objeto garantizar que las pruebas de hipótesis se llevan a cabo en condiciones correctas (nivel de escala, requisitos previos de la prueba, etc.). Lo mismo ocurre con la verificación gráfica de los modelos o la estimación de si los datos o los errores se distribuyen normalmente. Esto difiere de las explicaciones anteriores que sirven para probar el modelo lo mejor posible antes de aplicarlo a otro conjunto de datos en el futuro. Hoy en día, afortunadamente, se suelen aplicar procedimientos que trabajan *menos a ciegas*, sino que se aproximan a los datos en el sentido del modelo y revisan sucesivamente estos modelos en función de los conocimientos adquiridos. El resultado de tal modelización no es ciertamente de naturaleza confirmatoria, aunque el resultado suene plausible. *Siempre* requiere la aplicación a un nuevo conjunto de datos. Sin embargo, este argumento también se aplica a los análisis supuestamente confirmatorios, aunque tengan lugar en condiciones experimentales, ya que siempre existe la posibilidad de que una falsación en sí misma contenga errores y que dicha falsación se asuma

entonces erróneamente como verdadera. Esto demuestra que las técnicas de AED pueden integrarse perfectamente en los métodos estadísticos actuales, como el modelo y la prueba crítica de modelos. Los siguientes estudios de casos muestran posibles enfoques para obtener nuevos conocimientos a partir de los datos.

Así, AED nunca tiene como objetivo la confirmación, sino siempre la exploración. El AED nunca es problemático. Sin embargo, debe quedar claro qué afirmaciones pueden derivarse y cuáles no y documentar y comunicar esto al mundo exterior exactamente de la misma manera.

5.5 Casos prácticos de AED

Con la ayuda de varios estudios de casos empíricos y conjuntos de datos, se discutirá el uso de AED y -- si es posible – su potencial para los métodos mixtos. Probaremos técnicas típicas de AED para explorar de forma lúdica conjuntos de datos accesibles al público y datos de R. Además del código R, la atención se centra en los procesos de toma de decisiones subyacentes. Éstos se verbalizan con el mayor detalle posible en el sentido de pensar en voz alta para inspirar el propio enfoque. Resulta que los análisis no discurren con fluidez y elegancia por sí solos, sino que a veces hay desvíos y giros en U en callejones sin salida son necesarios para llegar a resultados razonables. Los casos prácticos son:

1. **Población en los Estados Federados** – Características de la población de la RFA de 2016 y 2017 respectivamente. Para 2016, encontramos para cada estado federal el número de habitantes, la superficie, los habitantes por km2, las defunciones y los nacimientos, así como la edad media. El modelo es Tukey (1977, p.148.) con datos comparables de EE.UU. (por estado) para el año 1959/1960, en el que el autor intentó comparar gráficamente la relación entre nacimientos y defunciones con la densidad de población. El resultado fue una estructura en forma de L – por tanto torcida, sesgada y desfavorable para el cálculo. Tras varias transformaciones de los datos Tukey tomó el logaritmo de la relación entre nacimientos y muertes y lo representó gráficamente con la mediana de edad por estado. El resultado fue que la nube de puntos de los datos estaba ahora alineada en una línea relativamente "recta", de modo que una simple regresión lineal podía seguir, o viceversa, que la edad mediana era la relación entre log(nacimientos/muertes) y la edad mediana. Los residuos del modelo lineal se distribuyeron discretamente alrededor de CERO. Reproducimos algo similar con datos de la población alemana, ya que el procedimiento ilustra muy bien el propósito de una transformación de datos no lineal.

- **Los datos suizos sobre fertilidad** – otro tema interesante de AED y un conjunto de datos elaborado por Tukey (Mosteller & Tukey, 1977, Proyecto "16P5", pp.549-551). Se dispone en R de los datos suizos de 1888 de 47 provincias francófonas sobre la fecundidad y sus vínculos con indicadores sociales como el trabajo (por ejemplo, "¿Se dedican los hombres a la agricultura?"), el porcentaje de catolicismo, la aptitud para el ejército, la educación y la mortalidad infantil, todos ellos disponibles en relación con la población. La fecundidad se obtiene a partir de una medida normalizada. Nos acercamos a este conjunto de datos sin ninguna hipótesis orientadora y vemos lo que surge.

- **Los datos de Iris** – recogidos por Edgar Anderson en 1936 (Anderson, 1935, 1936). Se trata de datos clásicos de la biología. Aquí tenemos que descubrir qué estructuras hay en los datos. El estadístico Ronald A. Fisher (1936) utilizó este conjunto de datos para demostrar el análisis discriminante lineal (LDA). Consta de 50 muestras cada una de tres especies diferentes (Iris setosa, Iris virginica, Iris versicolor). Se midieron cuatro propiedades por muestra (longitud y anchura del sépalo y del pétalo). Fisher utilizó el LDA para distinguir las especies entre sí (clasificación). Examinamos si lo consiguió, cómo puede hacerse visualmente sin análisis y cómo puede ser una revisión crítica del análisis con métodos AED.

2. **El hundimiento del Titanic** – con la supervivencia y las muertes en el gran desastre marítimo el 15. 04. 1912 en el Atlántico Norte. La cuestión de interés es qué características de los pasajeros eran favorables para la supervivencia, independientemente del comportamiento real de los individuos a bordo. Este conjunto de datos ha dado lugar a una plétora de publicaciones de carácter confirmatorio, para que cualquiera pueda obtener información adicional por su cuenta. Por ejemplo, podríamos investigar si es cierto que "mujeres y niños primero" era válido – o si la riqueza ofrecía una ventaja para la supervivencia.

- **Liderazgo en el sector educativo español** – a partir de datos de entrevistas cualitativas y y las capacidades de codificación temática de AQUAD 7 (Huber & Gürtler, 2012) tratamos de mostrar cómo es posible encontrar patrones en los datos. El contexto es el sector educativo español y el contenido es el tema del liderazgo. El objetivo es comprender de qué tratan realmente las entrevistas en profundidad y qué dicen y cómo se relacionan con un estudio de cuestionario más amplio realizado con anterioridad.

- **Un experimento quiropráctico** – se trata de un diseño deductivo de comprobar una hipótesis en el contexto del trabajo quiropráctico (Wiper, 2017). Sin embargo, en lugar de basarse en una lógica de significación puramente probatoria, se ha utilizado el AED para desarrollar hallazgos más profundos. Los resultados ofrecen una base interesante para planificar y llevar a cabo una replicación del estudio y aplicar correctamente los resultados de un modelo jerárquico-lineal.

- **El potencial de reintegración en la terapia de la adicción** – el último estudio de caso compara los resultados de los análisis cualitativos de casos individuales cualitativos con los de una simple categorización cuantitativa del mismo conjunto de datos para evaluar el potencial de recuperación de antiguos adictos en el contexto de un estudio de catamnesis en Suiza (Gürtler, Studer & Scholz, 2012).

En cada caso, elaboramos el AED hasta el punto dónde queda claro el funcionamiento básico de sus técnicas y los comandos R asociados, es decir, ninguno de los análisis se lleva hasta el final. Sin embargo, se dan pistas sobre dónde se puede utilizar R para otros análisis y cómo se puede encontrar más información en Internet y en la literatura pertinente. Los lectores interesados pueden simplemente continuar con los análisis y ver qué más pueden descubrir en los datos o incluso cuestionar y contrastar críticamente nuestras modestas conclusiones.

5.5.1 La población en una comparación de los estados federales de Alemania

No seguimos todo el proceso de Tukey (1977, p.148.) y nos limitamos sólo a los habitantes por estado federal de Alemania. Los datos brutos del archivo `German_states_population.tab` ya contienen toda la información necesaria para reproducir completamente los pasos de Tukey descritos anteriormente. Por lo tanto, examinamos primero la tabla de datos. Contiene datos del año 2016, que están disponibles en diversas fuentes gratuitas. La tabla 5.1 explica el conjunto de datos y sus variables (`ptII_quan_EDA_case-estados-alemanes-poblacion.r`).

```
# Población de los Estados Federales de Alemania
# el 31 de diciembre 2017 (en 1.000)
einw2017 <- c(17912,12997,11023,7963,6243,4081,4074,3613,
2890,2504,2223,2151,1831,1611,994,681)
bland <- c("Nordrhein-Westfalen","Bayern","Baden-Württemberg", "Niedersachsen",
"Hessen","Sachsen","Rheinland-Pfalz", "Berlin","Schlwesig-Holstein", "Brandenburg",
"Sachsen-Anhalt","Thüringen","Hamburg", "Mecklenburg-Vorpommern","Saarland","Bremen")
eastwest <- c("W","W","W","W","W","E","W","B",
"W","E","E","E","W","E","W","W")
# 2016
```

```
# km^2 in 1000
area2016 <- c(34097,70551,35751,47634,21114,18415,19853,
891,15799,29485,20446,16172,755,23180,2569,419)
# population
einw2016 <- c(17877808,12887133,10915756,7936142,6194630,4083317,
           4059428,3547431,2870320,2489737,2240861,2164421,
           1798923,1611518,996124,675121)
# people per km^2
ppkm2.2016 <- c(524,183,306,167,294,221,205,4012,182,84,109,133,2397,69,388,1617)
# deaths
death2016 <- c(202250,129552,106630,92368,64081,53330,45863,34050,
33879,30790,31453,28312,17267,20445,12897,7732)
# births
births2016 <- c(173274,125686,107479,75215,60731,37940,37518,41086,
           25420,20934,18092,18475,21480,13442,8215,7136)
# mean age
mage2016 <- c(44,43.6,43.3,44.4,43.7,46.7,44.6,42.6,
           45,46.9,47.5,47,42.1,46.7,46.1,43.6)
poptab <- data.frame(einw2016,area2016,ppkm2.2016,
           death2016,births2016,mage2016)
rownames(poptab) <- bland
colnames(poptab) <- c("population","km^2","ppkm^2","death","birth","age (mean)")
write.table(poptab,file="German_states_population.tab",
           sep="\t", row.names=TRUE, col.names=TRUE)
rm(poptab)
poptab <- read.table(file="German_states_population.tab",header=TRUE, sep="\t")
poptab
```

La tabla se ordena según el número de habitantes. Es posible realizar otra ordenación con `order()`. Ordenamos según el estado federal con la tasa de natalidad absoluta más baja:

```
# show tab R-Code
poptab
# order along minimum births
poptab[order(poptab$birth),]
```

Tab. 5.1: *Habitantes de los estados federales — Variables*

Variable	*Valor*	*Significado*	*Tipo*
einw2017	Habitantes en 1000	Inhab. 2017 del estado	numeric
bland	Nombre	Estados federales	string
eastwest	W(est), E(ast) [Oeste/Este]	Asignación antes de la reunificación BRD vs. DDR	logic
area2016	km^2	Superficie del estado federal	integer
einw2016	Habitantes en 1000	Habitantes 2016 del estado	integer
$ppkm^2$2016	Personas	Habitantes per km^2	integer
death2016	Personas	muertos per estado	integer
birth2016	Personas	Nacimientos per estado	integer
mage2016	Edad	Edad media per estado	numeric

Para la demostración de la transformación de datos tomamos el número de habitantes por estado federal del año 2017, que está contenido en el vector *einw2017*. Primero observamos el diagrama de barras (véase la Fig. 5.6).

```
# barplot
par(mar=c(7,4,2,2)+0.2)
bpl <- barplot(einw2017, col=rainbow(length(einw2017)),
border="blue", xlab="",main="Population German Federal Lands")
text(bpl, par("usr")[3], srt=60, adj=1,xpd=TRUE, labels=bland, cex=0.65)
```

Podríamos preguntarnos ahora si existe una gran diferencia entre el número de habitantes de los Estados federados orientales (antigua RDA) y los occidentales (antiguos Estados federados de la RFA, véase la Fig. 5.7). Asignamos Berlín al Este y al Oeste.

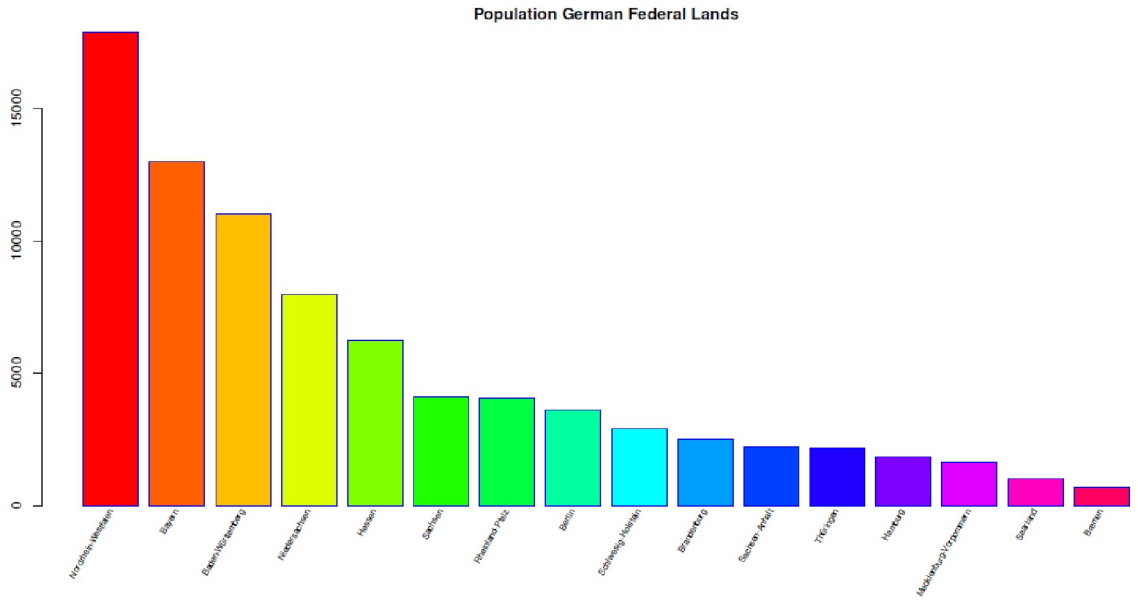


Figura 5.6. *Habitantes pro estado federal*

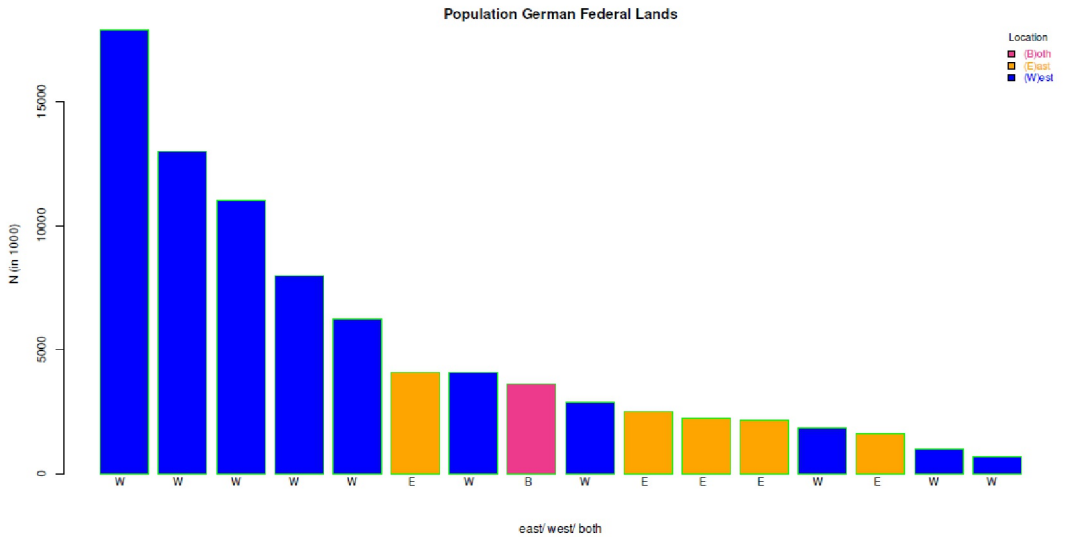


Figura 5.7. *Habitantes pro estado federal (Comparación Oeste/Este)*

```
# plot according to east/west/both
par(mar=c(5,4,2,2)+0.2)
fac <- 1.9
bcolo <- c("violetred2","orange","blue")
```

```
bpl2 <- barplot(einw2017,col=bcolo[as.factor(eastwest)],
            border="green",main="Population German Federal Lands",
            xlab="east/ west/ both",ylab="N (in 1000)",
            legend.text=c("(B)oth","(E)ast","(W)est"),
            args.legend=list(x="topright",title="Location",
            cex=.85, bty="n",fill=bcolo, title.col="black",
            text.col=c("violetred2","orange","blue")))
text(bpl2, par("usr")[3]*fac, srt=0, adj=1,
            xpd=TRUE, labels=eastwest, cex=.9)
```

Esto favorece claramente a los Estados occidentales. Ahora nos interesa una visión general de la distribución de los datos. Por desgracia, `fivenum()` no proporciona ninguna etiqueta para los valores:

```
> # Tukey's fivenum descriptive statistics
> fivenum(einw2017)
[1] 681.0 1991.0 3251.5 7103.0 17912.0
> # no labes, better use one's own function
```

En consecuencia, escribimos una función corta de R, es decir, un wrapper llamado `fivenum.wn()`, que hace el y la llamamos inmediatamente.

```
> fivenum.wn(einw2017)
Min 1st Qu. Median 3rd Qu. Max
681.0 1991.0 3251.5 7103.0 17912.0
> mean(einw2017)
[1] 5174.438
```

Se observa que la mediana está más cerca del mínimo que del máximo y que la media y la mediana son claramente diferentes. Para verlo más de cerca, trazamos los datos como un simple diagrama de dispersión en orden descendente (véase la Fig. 5.8), que en este caso corresponde al diagrama de barras anterior.

```
# plot population for each German Federal Land
bl <- 1:length(bland)
einw2017.lm <- lm(einw2017~bl)
einw2017.log.lm <- lm(log(einw2017) ~ bl)
par(mfrow=c(1,2))
plot(bl,einw2017, bty="n", pre.plot=grid(),
main="Population per German Federal Land", pch=21, bg="darkred",
            col="blue", ylab="population", xlab="German Federal Land")
```

y añadimos una línea de regresión (véase la Fig. 5.8) estimando una regresión lineal simple con `lm()` y trazándola con `abline()`.

```
abline(einw2017.lm, col="violetred2", lty=2)
```

Todavía no parece muy recto. No es completamente una estructura en forma de L como la de Tukey, pero sigue teniendo mucha forma de arco y de L. Es importante tener en cuenta que, para el gráfico, hemos clasificado los Estados federales según el número de habitantes, de modo que el eje X se corresponde en última instancia con el eje Y, salvo que en el eje X las distancias entre los puntos son iguales, mientras que no ocurre lo mismo con el eje Y (= número de habitantes). El objetivo de la transformación de los datos sería igualar las distancias relativas entre el eje X y el eje Y, de modo que una línea de regresión recta cubra relativamente bien todos los puntos. Por lo tanto, ahora logaritmizamos el número de habitantes y repetimos el gráfico (véase la Fig. 5.8).

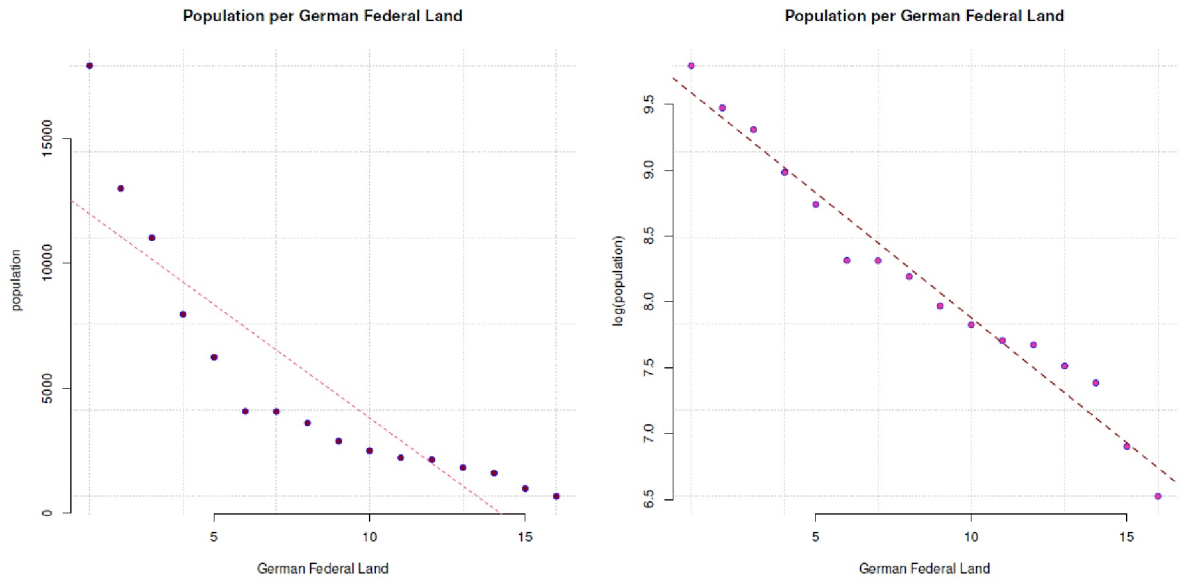


Figura 5.8. *Número de habitantes por estado federal (sin logaritmizar o logaritmizado)*

Al mismo tiempo estimamos de nuevo el modelo lineal mediante regresión y trazamos la recta de regresión resultante.

```
# plot log(population) for each German Federal Land
plot(bl,log(einw2017), bty="n", pre.plot=grid(),
           main="Population per German Federal Land",
           pch=21, bg="violetred2", col="blue",
           ylab="log(population)", xlab="German Federal Land")
abline(einw2017.log.lm, col="darkred", lty=2, lwd=1.8)
```

Ahora hay que trazar los dos modelos lineales (véase la Fig. 5.9) con fines de diagnóstico. El gráfico inferior contiene los diagnósticos con logaritmación (puntos azules).

```
par(mfrow=c(2,4))
plot(einw2017.lm, col="violetred3", pch=21, bg="darkred")
plot(einw2017.log.lm, col="azul", pch=21, bg="azul")
```

Obviamente, los residuos parecen mejor distribuidos en el curso de la transformación `log()` y el gráfico de los residuos frente a la distribución normal muestra desviaciones significativamente menores de la expectativa que sin la transformación `log()`, especialmente en los extremos. Del mismo modo, el gráfico de los valores ajustados frente a la raíz de los residuos normalizados parece mucho menos aleatoria. Observemos los resultados de la estimación del modelo, esta impresión se confirma. Ignoramos aquí deliberadamente cualquier referencia a *p*-valores y significancias en la salida y por lo tanto recurrimos a `display()` del paquete R `arm` para la salida.

```
# check linear models R-Code
einw2017.lm.sum <- display(einw2017.lm)
einw2017.log.lm.sum <- display(einw2017.log.lm)
t.wo.log <- with(einw2017.lm.sum, coef/se)
t.w.log <- with(einw2017.log.lm.sum, coef/se)
```

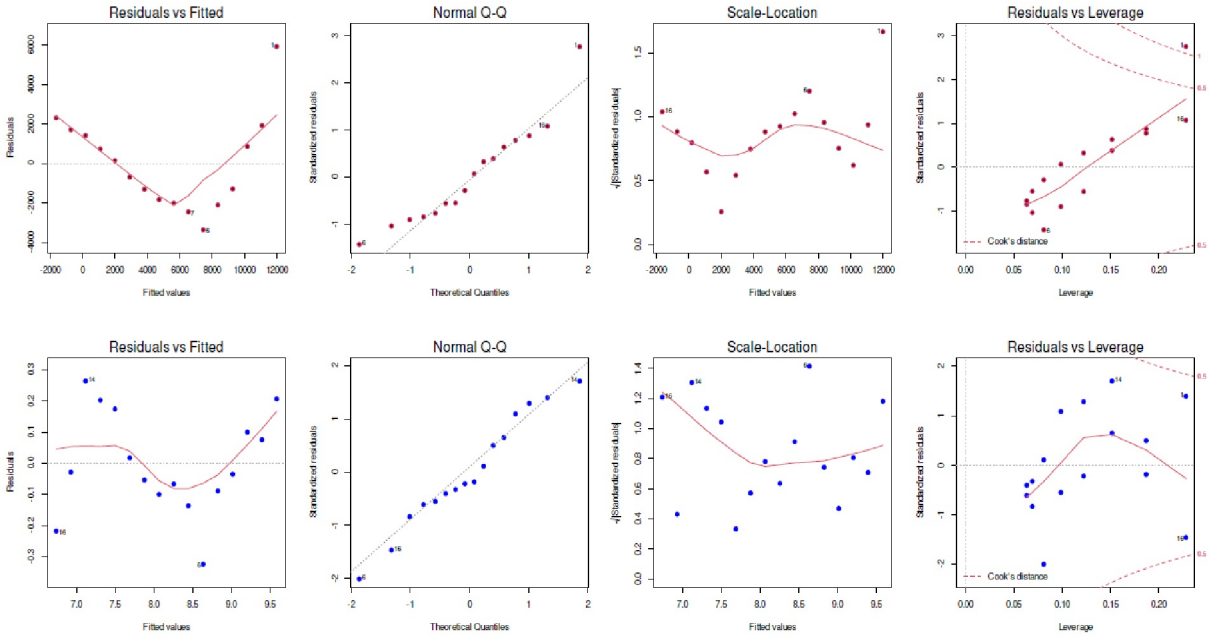


Figura 5.9. *Número de habitantes por estado federal (diagnóstico de regresión, sin logaritmizar o logaritmizado)*

El error estándar de la β-ponderación ha disminuido claramente debido a la transformación `log()`, lo que puede observarse en el triple valor *t*, cociente de la estimación de la ponderación y su error estándar.

```
> t.wo.log
(Intercept) bl
10.05217    -6.84190
> t.w.log
(Intercept) bl
110.60374   -20.75805
> t.w.log/t.wo.log
(Intercept) bl
11.002966    3.033961
```

El cociente de los valores *t* entre el modelo con y sin `log()` muestra una diferencia de factor 3 a favor del modelo con `log()`. El R^2 de ajuste del modelo (con y sin ajuste por el número de predictores en el modelo) también ha mostrado un aumento sustancial de

```
> # compare R^2
> 1-einw2017.log.lm.sum$r.squared/einw2017.lm.sum$r.squared
[1] -0.2581921
```

lo que supone un aumento del 25.8%. El gráfico de dispersión anterior (véase la Fig. 5.8) muestra que una línea de regresión recta describe bastante bien los datos, de modo que las distancias en el eje Y corresponden ahora aproximadamente a las del eje X. Por supuesto, esta transformación debe tenerse en cuenta en las interpretaciones posteriores del contenido. Ya no estamos en la escala original, sino en el logaritmo de la escala original. Pero incluso esto puede tener ventajas para una interpretación, como muestran Gelman y Hill (2007). Podríamos – como sugiere a menudo Tukey (1977) – también podríamos trabajar con raíces cuadradas, lo que, sin embargo, parecería menos favorable y está en algún lugar entre ninguna y la transformación `log()`:

```
> # do sqrt and compare to the other solutions
> display(lm(sqrt(einw2017) ~ bl))
lm(formula = sqrt(einw2017) ~ bl)
```

```
            coef.est coef.se
(Intercept) 117.67    5.36
bl           -6.15    0.55
---
n = 16, k = 2
residual sd = 10.22, R-Squared = 0.90
```

No podemos explicar nada más allá de eso, porque ya estaba claro de antemano que los estados federales difieren. La transformación `log()` no añadió información nueva. De todos modos, el objetivo de esta demostración no era "explicar más", sino mostrar que una transformación de datos no lineal adecuada a cada caso puede dar lugar a resultados sorprendentes que pueden facilitar mucho los cálculos posteriores. En este caso, se trataría de la aplicabilidad de modelos lineales frente a modelos no lineales o modelos lineales con términos predictores cuadráticos o superiores (= regresión polinómica). En otras palabras, una *simplificación del proceso de trabajo*.

Tarea 5.1: Transformaciones de datos

Utilizando los datos de 2016, realiza las transformaciones de datos y los cálculos que sugiere Tukey. ¿Llegas a un resultado tan favorable como el del autor o qué más tendrías que hacer? Podrías, por ejemplo, eliminar datos para fijarte sólo en las secciones y explicar los datos eliminados de otra manera o buscar diferentes transformaciones de los datos, por ejemplo, para explicar la proporción de nacimientos frente a defunciones según la densidad de población o la edad media (lamentablemente, no disponíamos de datos sobre la edad media de los estados federados en 2016). Dependiendo de su interés, puede formar otras proporciones dentro de los datos o utilizar los enlaces de Internet para encontrar datos adicionales para explorar con más detalle.

5.5.2 Fecundidad y fertilidad

En este conjunto de datos decidimos preguntarnos si se puede establecer una posible conexión entre el catolicismo y la fertilidad (lema bíblico: "Sed fecundos y multiplicaos", Génesis 9.7) y de qué manera. En el caso de los actuales territorios suizos a finales del siglo XVIII (recopilado por Francine Vanderwalle), una baja proporción de catolicismo presumiblemente significaba directamente una proporción correspondientemente alta de reformismo/protestantismo o calvinismo (= denominación de las iglesias reformadas en Suiza o del sistema teológico de Juan Calvino, 1509-1564). Por tanto, estamos examinando una relación diametral entre catolicismo y protestantismo. En primer lugar, tenemos que examinar el conjunto de datos y describirlo descriptivamente (`ptII_quan_EDA_caso_Suisse-fertilidad.r`):

```
?swiss
head(swiss)
tail(swiss)
describes(swiss)
```

Recibimos una descripción abreviada con

```
> t(apply(swiss,2,summary))
                  Min. 1st Qu. Median Mean   3rd Qu. Max.
Fertility         35.00 64.700 70.40 70.14255 78.450  92.5
Agriculture        1.20 35.900 54.10 50.65957 67.650  89.7
Examination        3.00 12.000 16.00 16.48936 22.000  37.0
Education          1.00  6.000  8.00 10.97872 12.000  53.0
Catholic           2.15  5.195 15.14 41.14383 93.125 100.0
Infant.Mortality  10.80 18.150 20.00 19.94255 21.700  26.6
```

No faltan valores. La página de ayuda del conjunto de datos menciona n = 47 observaciones sobre k = 6 variables. Todas las variables, excepto la fecundidad, contienen proporciones de población (escaladas de 0; : : ; 100). La fecundidad es una medida normalizada (véase también Oce of PopuFrancine Vanderwallelation Research, 2019). Las demás variables sólo tienen un interés marginal y no desempeñan ningún otro papel. No se trata de crear un modelo lineal con todas las variables como predictores y optimizarlo para explicar la fecundidad. El interés se concentra en la relación entre fecundidad y catolicismo. Un diagrama de tallo y hoja señala una para la fecundidad y una distribución de dos colas para el catolicismo.

```
> # could be an U-shape
> stem(swiss$Fertility)
The decimal point is 1 digit(s) to the right of the |
3 | 5
4 | 35
5 | 46778
6 | 124455556678899
7 | 01223346677899
8 | 0233467
9 | 223

> stem(swiss$Catholic)
The decimal point is 1 digit(s) to the right of the |
0 | 2223333445555566899
1 | 0124578
2 | 4
3 | 4
4 | 2
5 | 08
6 |
7 |
8 | 5
9 | 113377889999
10 | 000
```

Así que parece que hay un SÍ o un NO católico en estas áreas y nada en medio. Desde el punto de vista histórico, es probable que esto sea incluso correcto. Un diagrama de dispersión (véase la Fig. 5.10) de ambas variables con `scaplot.cont()` para ilustrar el grado continuo de la variable *Católico* da una estimación de la influencia de los distintos grados de Catolicismo

```
scaplot.cont(x=swiss$Catholic, y=swiss$Fertility, R-Code
TITLE="Swiss data on Fertility from 1888",
```

permite hacer varias suposiciones. Parece haber diferencias entre el alto y el bajo catolicismo en términos de fertilidad. Esto es motivo suficiente para crear dos niveles exclusivos a partir de cada una de las dos variables y para examinarlos combinatoriamente en busca de correlaciones lineales. El primer paso es la correlación global, pero debemos suponer que es engañosa (véase la Fig. 5.10).

```
> # correlations fertility and catholic index
> # general r = .46
> with(swiss, cor(Fertility, Catholic))
[1] 0.4636847
```

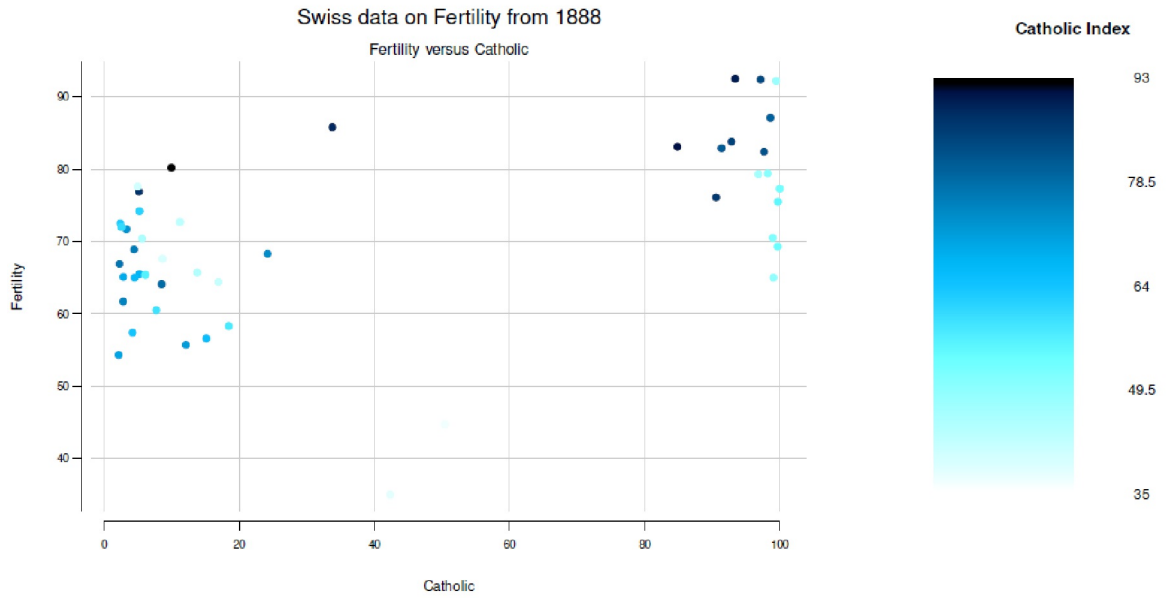


Figura 5.10. *Datos suizos (fecundidad frente a catolicismo, diagrama de dispersión)*

Con r = 0.46 parece existir una correlación. Como es habitual con un AED, la prueba de significación no desempeña ningún papel. Si sólo se divide la variable Católicos o sólo Fecundidad, surge una imagen diferente. Utilizamos la Figura 5.10 como guía para fijar el punto de corte. A continuación, dividimos y combinamos las dos variables Fecundidad y Católica en cuatro áreas separadas y calculamos las correlaciones respectivas.

```
# add new variables Fertility and Catholic split at 50%
# correlations for F & C hi versus lo
nk <- 2
swiss2 <- data.frame(swiss,
           F.hi=swiss$Fertility > 50,
           C.hi=swiss$Catholic > 50)
swiss2
F.hilo <- factor(swiss2$F.hi, labels=c("F.lo","F.hi"))
C.hilo <- factor(swiss2$C.hi, labels=c("C.lo","C.hi"))
swiss2$FC <- factor(paste(F.hilo, C.hilo, sep=":"))
head(swiss2)
r <- NULL
for(i in levels(swiss2$FC))
{
r[i] <- round(with(swiss2[swiss2$FC == i,],
           cor(Fertility, Catholic)),nk)
cat(paste(i, " --- r = ", r[i],"\n",sep=""))
}
```

Los rangos para r = NA y r = -1 necesitan un examen más detallado, porque la aparición de tales valores siempre tiene una razón, que rara vez se encuentra en la propia pregunta. Resulta que el primer caso sólo contiene un dato y, por tanto, no permite ninguna correlación (¿con qué?). El otro caso se basa en n = 2 casos y no tiene mucho más sentido.

```
> #r=-1
> subset(swiss2, Fertility < 50 & Catholic > 50)
           Fertility Agriculture Examination Education Catholic
Rive Droite 44.7       46.6        16          29        50.43
Rive Gauche 42.8       27.7        22          29        58.33
```

```
           Infant.Mortality F.hi  C.hi FC
Rive Droite 18.2            FALSE TRUE F.lo:C.hi
Rive Gauche 19.3            FALSE TRUE F.lo:C.hi
> #r=NA
> subset(swiss2, Fertility < 50 & Catholic < 50)
           Fertility Agriculture Examination Education Catholic
V. De Geneve   35      1.2         37          53        42.34
           Infant.Mortality F.hi  C.hi  FC
V. De Geneve   18           FALSE FALSE F.lo:C.lo
```

Gráficamente tiene este aspecto (véase la Fig. 5.11):

```
# two color split R-Code
# have a closer look on the various possibilities
plot(swiss$Catholic, swiss$Fertility, pch=21,
bg=c("violetred2","blue")[(swiss$Catholic > 50)+1],
bty="n", main="Swiss data (1888)")
abline(v=50, lty=2, col="red")
abline(h=50, lty=2, col="steelblue")
r.xy <- data.frame(x=c(70,20,70,20), y=c(90,90,40,40),
colo=c("violetred2","blue","olivedrab","brown"))
for(i in 1:length(r))
{
text(r.xy[i,c("x","y")],paste(names(r)[i]," = ",r[i],sep=""),
col=as.character(r.xy[i,"colo"]))
}
```

En consecuencia, parece justificado mantener el punto de vista de los distintos casos y examinar cada uno por separado. Una nota al margen: obviamente, las zonas fronterizas de católicos sufren sobre todo efectos *techo y suelo* (Krauth, 1995), que pueden explicarse por el contexto histórico (= religiosidad en Suiza a finales del siglo XIX, presumiblemente zonas rurales) y tendrían que considerarse por separado. La zona de *baja fecundidad* podría considerarse junto con la variable `Catholic`, ya que los valores de `Catholic` se sitúan en un rango estrecho y medio entre el 42%-58% y gráficamente (véase la Fig. 5.11) no se puede justificar una separación, al contrario que en las zonas extremas de `Catholic`.

```
> # Catholic - only lo Fertility
> range(subset(swiss2, Fertility < 50)$Catholic)
[1] 42.34 58.33
```

La extensión de las agrupaciones se muestra en el boxplot de la Figura 5.12.

```
TITLE <- "Swiss data on Fertility from 1888"
SUB <- "Fertility high/low vs. Catholic high/low"
boxplot(Fertility ~ FC, data=swiss2, notch=TRUE,
col=c("olivedrab3","steelblue","orange","yellow"),
boxwex=1, lex.order=TRUE, varwidth=TRUE,
outline=TRUE, frame.plot=FALSE, border=2:5)
rug(swiss$Fertility, side=2, col="magenta3")
mtext(TITLE, 3, line=2, cex=1.5)
mtext(SUB, 3, line=.6, cex=1.1)
```

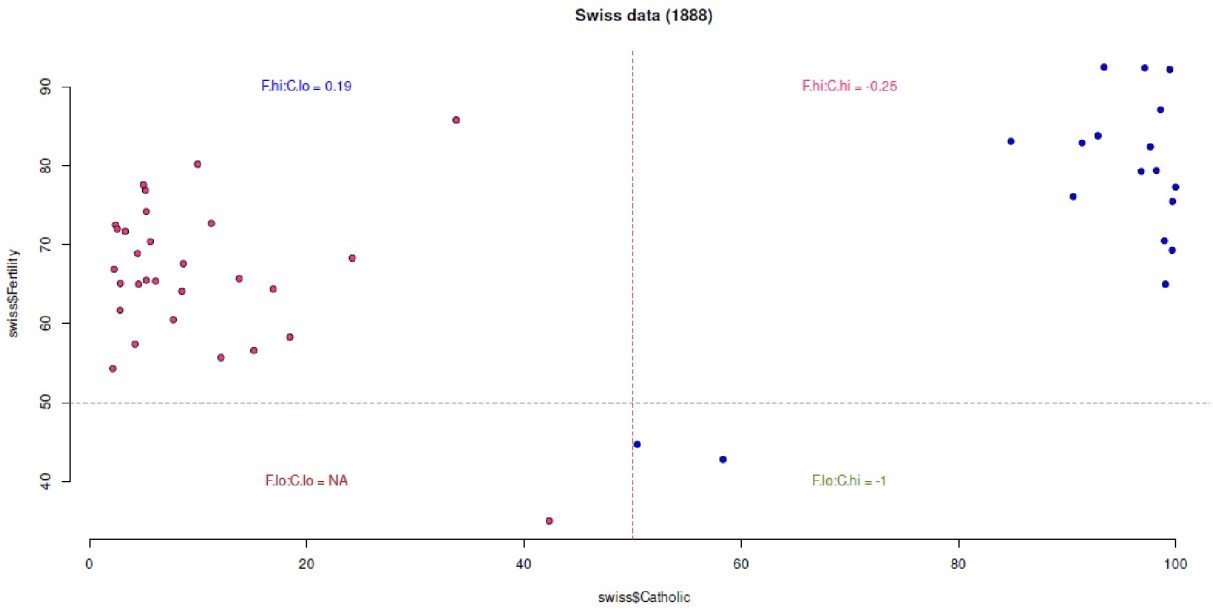


Figura 5.11. *Datos suizos (fecundidad y catolicismo, correlaciones)*

Por lo tanto, tiene sentido combinar los valores del catolicismo para la baja fecundidad y repetir el boxplot.

```
# combine factor levels for F
levels(swiss2$FC)
swiss2$FC.adj <- combineLevels(swiss2$FC,
                 levs=c("F.lo:C.lo","F.lo:C.hi"),
                 newLabel="F.lo.C")
# check
swiss2[,c("FC","FC.adj")]
# boxplot
SUB <- "Fertility vs. Catholic high/low"
boxplot(Fertility ~ FC.adj, data=swiss2, notch=TRUE,
        col=c("olivedrab3","steelblue","orange"),
        boxwex=1, lex.order=TRUE, varwidth=TRUE,
        outline=TRUE, frame.plot=FALSE, border=2:5)
rug(swiss$Fertility, side=2, col="magenta3")
mtext(TITLE, 3, line=2, cex=1.5)
mtext(SUB, 3, line=.6, cex=1.1)
```

Ahora hay tres áreas cuyos organismos de datos apenas parecen solaparse. Llegados a este punto, se podría estimar un modelo lineal y calcular los coeficientes mediante simulación. Antes de eso, es importante echar un vistazo a las estadísticas descriptivas de las tres zonas respecta a la fecundidad.

El modelo y los parámetros estimados mediante simulación pueden interpretarse ahora en términos de contenido. ¿Es cierta la afirmación bíblica "Creced y multiplicaos", que en este caso se refiere al catolicismo (véase más arriba)? Para estos datos – SÍ, pero la afirmación es igualmente cierta para el muy bajo catolicismo y, por tanto, no es realmente significativa a nivel mundial. Pero si es cierto que, en relación con la muestra, bajo catolicismo es sinónimo de alto puritanismo y calvinismo en lugar de no religión – lo que habría que determinar –, entonces queda el contexto bíblico pero no el católico. Pero esto no es ni una prueba ni una reconstrucción de relaciones causa-efecto. De hecho, éstas sólo se aclaran cuando se entiende cómo se relaciona una baja fecundidad con una religiosidad presumiblemente más baja, y si o cómo otros factores influyentes de forma complementaria afectan directa o indirectamente a ambas variables (por ejemplo, el estatus socioeconómico, la atención médica, el acceso a los recursos médicos, etc.). Del mismo modo, sigue

sin estar claro que un catolicismo muy elevado siga teniendo una fecundidad significativamente mayor que, posiblemente, una alta prevalencia de puritanismo/calvinismo. Esto podría dar lugar a un debate fascinante en el que no se considere necesariamente a la religión como la causa, sino que se analicen las variables mediadoras.

```
> # descriptive
> do.call("cbind",with(swiss2, tapply(Fertility, FC.adj, fivenum2)))
               F.hi:C.hi  F.hi:C.lo   F.lo.C
minimum        65.00      54.30      35.00
lower-hinge    75.80      62.90      38.90
median         80.90      66.30      42.80
upper-hinge    85.45      72.25      43.75
maximum        92.50      85.80      44.70
> # linear model
> seed <- 9876
> table(swiss2$FC.adj)
F.hi:C.hi F.hi:C.lo F.lo.C
16 28 3
> display(swiss2.lm <- lm(Fertility ~ FC.adj, data=swiss2))
lm(formula = Fertility ~ FC.adj, data = swiss2)
                  coef.est coef.se
(Intercept)        80.55    1.94
FC.adjF.hi:C.lo   -13.21    2.43
FC.adjF.lo.C      -39.72    4.88
---
n = 47, k = 3
residual sd = 7.76, R-Squared = 0.63
> set.seed(seed)
> swiss2.sim <- sim(swiss2.lm, n.sims=1000)
> str(swiss2.sim)
Formal class 'sim' [package "arm"] with 2 slots
..@ coef : num [1:1000, 1:3] 79.2 78.8 84.5 82.9 84.2 ...
.. ..- attr(*, "dimnames")=List of 2
.. .. ..$ : NULL
.. .. ..$ : chr [1:3] "(Intercept)" "FC.adjF.hi:C.lo" "FC.adjF.lo.C"
..@ sigma: num [1:1000] 7.07 7.83 7.34 7.94 9.14 ...
> apply(coef(swiss2.sim), 2, quantile)
     (Intercept)   FC.adjF.hi:C.lo FC.adjF.lo.C
0%      71.94689    -23.11120      -57.54758
25%     79.27922    -14.82670      -43.71952
50%     80.55944    -13.17130      -39.78487
75%     81.92744    -11.57689      -36.42801
100%    86.47046     -3.94837      -24.07810
> quantile(sigma.hat(swiss2.sim))
0%       25%      50%      75%      100%
5.941308 7.264820 7.779556 8.344946 11.697011
```

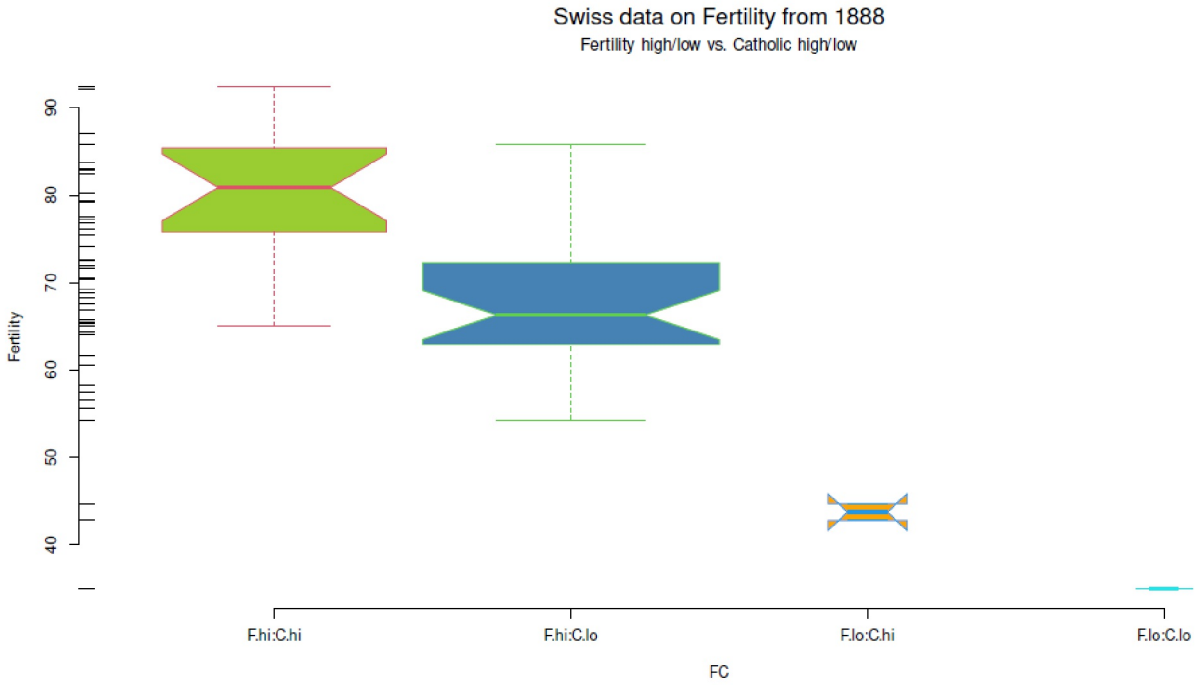


Figura 5.12. *Datos suizos (fecundidad y catolicismo, boxplots)*

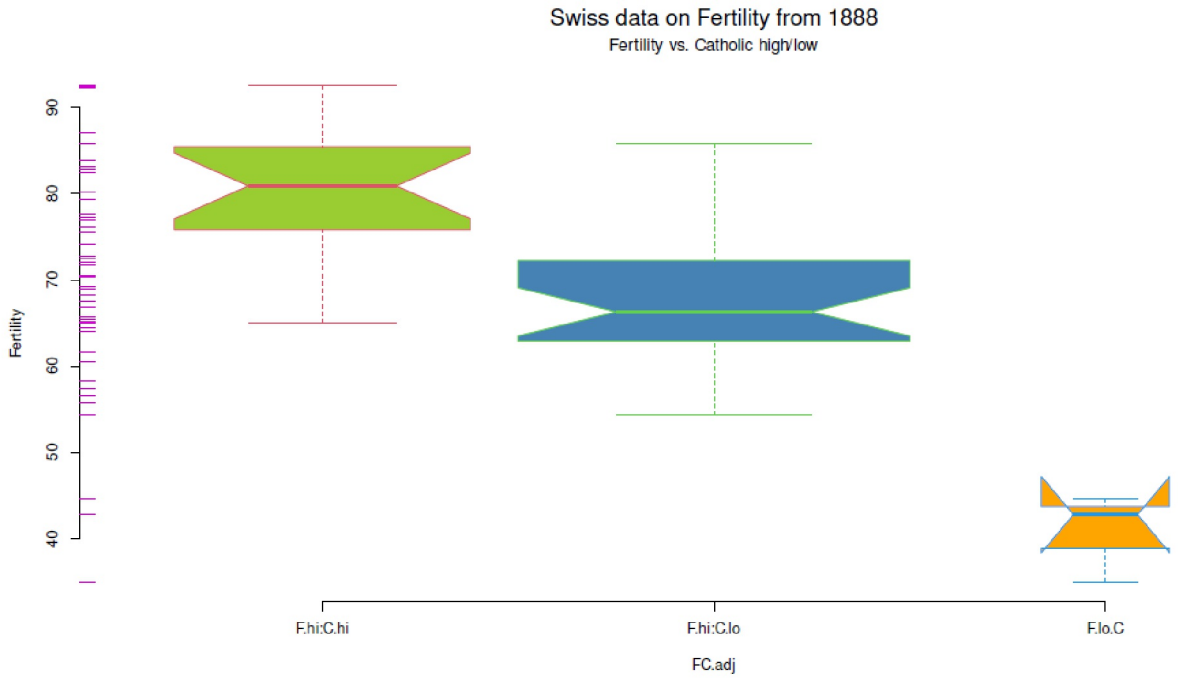


Figura 5.13. *Datos suizos (fecundidad y catolicismo, boxplots combinados)*

Lo que hemos encontrado a través de la AED son tres áreas distintas:

1. Una fecundidad generalmente baja, independientemente de la religión católica, y posiblemente característica de zonas con bajos niveles de religiosidad cristiana.

2. Muy alta fecundidad y muy alto catolicismo.
3. Alta fecundidad y muy bajo catolicismo, posiblemente alto puritanismo.

Un `smoothScatterplot()` optimizado para la borrosidad con una extensión basada en la sugerencia de por Arenburg (2016) también sugiere la separación de estas tres zonas (véase la Fig. 5.14).

Figura 5.14. *Datos suizos (fecundidad y catolicismo, smooth scatterplot)*

```
par(mar=c(5,4,4,6) + .1) R-Code
with(swiss, smoothScatter(Fertility, Catholic, nrpoints=0,
postPlotHook=fudgeit,
colramp=colorRampPalette(c("white", "darkorange"))))
title("")
SUB <- "Fertility vs. Catholic index"
mtext(TITLE, 3, line=2, cex=1.5)
mtext(SUB, 3, line=.6, cex=1.1)
```

El modelo lineal anterior es un indicio de que esta subdivisión tiene sentido y debe tomarse en serio. A partir de los conocimientos adquiridos surge ahora la pregunta: "¿Qué constituye una baja fecundidad? – ¿además de que el catolicismo es sólo moderado y no extremo? La tarea ahora es explorar esto e integrar las demás variables en el debate y, sobre todo, analizar el contexto histórico, por ejemplo, qué papel desempeñaban las religiones en aquella época. Así pues, la baja fecundidad podría ser objeto de un estudio cualitativo en profundidad con el fin de desvelar las verdaderas razones. Los posibles puntos de partida para encontrar interrelaciones entre las variables se muestran en los gráficos finales de las figuras 5.15 y 5.16.

```
# actual pairs plots
pairs(~ Fertility + Agriculture + Examination + Education +
           Catholic + Infant.Mortality + FC.adj, data=swiss2,
           diag.panel=panel.hist, upper.panel=panel.smooth,
           lower.panel=panel.cor, pch=21,
           bg=c("violetred2","skyblue")[unclass(swiss2$Catholic >50)+1],
           main="Swiss data (high vs. low <-> Fertility vs. Catholic)")
corrgram(swiss, order=TRUE,
           lower.panel=corrgram::panel.ellipse,
           upper.panel=corrgram::panel.cor,
           diag.panel=corrgram::panel.density,
           text.panel=corrgram::panel.txt,
           main="Swiss data",
```

```
            col.regions=colorRampPalette(c("darkgoldenrod4",
"burlywood1","darkkhaki", "darkgreen"))
)
```

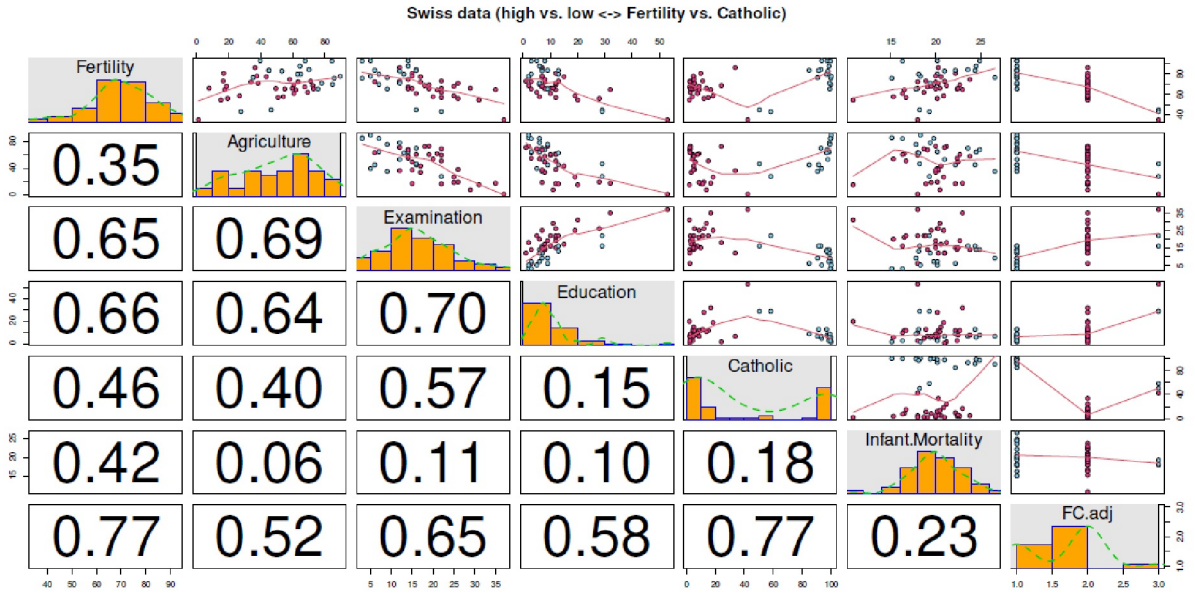


Figura 5.15. *Datos suizos (fecundidad y catolicismo, pairs plot)*

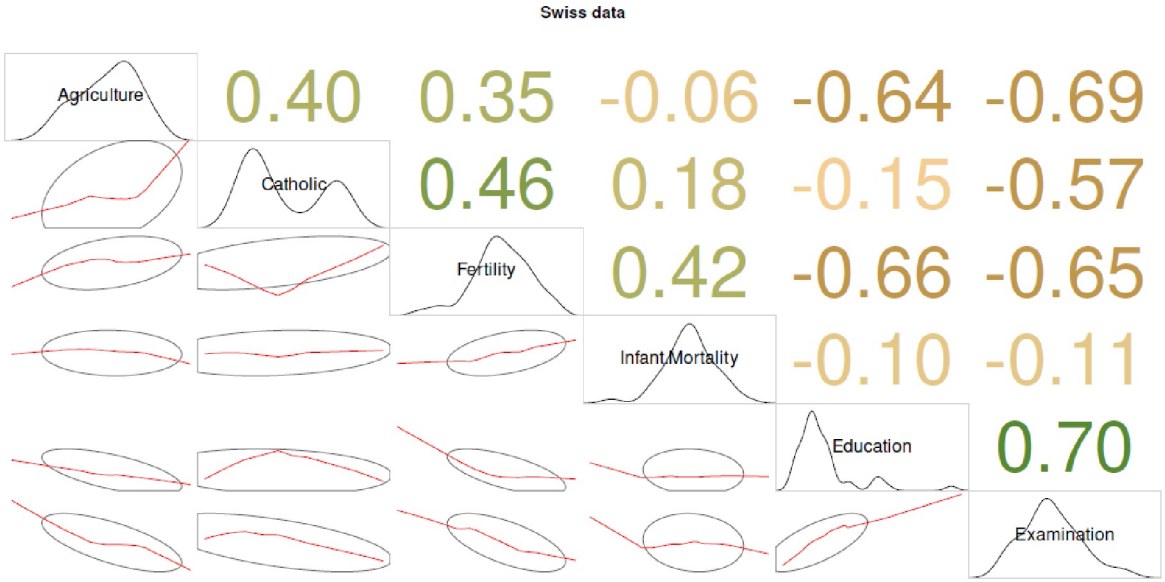


Figura 5.16. *Datos suizos (fecundidad y catolicismo, correlograma)*

Con esto concluye este AED para datos suizos. A continuación se presenta un AED en el contexto de los procesos de clasificación biológica.

5.5.3 Distinción de especies en biología

Los datos Iris de Edgar Anderson (1935, 1936) son uno de los conjuntos de datos clásicos de la estadística. Lo utilizamos para comprender cómo se puede clasificar gráficamente de antemano el enfoque de Ronald A. Fisher. A continuación reproducimos, sin parámetros estadísticos, el análisis exploratorio de Fisher, que no tiene carácter confirmatorio, es decir, sería si probara su análisis en un nuevo conjunto de datos y lo revisara críticamente. El punto de partida determinamos – como siempre – con una mirada a los datos brutos y estadísticas descriptivas robustas sensu Tukey con `fivenum.wn()`. Para el posterior análisis discriminante lineal seguimos necesitando la biblioteca `MASS` (`ptII_quan_EDA_case_Anderson_iris-especies-en-biologia.r`).

```
require(MASS)
# iris data set
# by Anderson and made famous by Fisher
data(iris)
head(iris)
tail(iris)
# lda iris data
di <- dim(iris)
apply(iris[,-5],2,fivenum.wn)
table(iris$Species)
```

Así que estamos tratando con m = 4 variables. Por lo tanto, un gráfico con `pairs()` podría ayudar.

```
# no color
pairs(iris[1:4], main="Iris data set (from Edgar Anderson)",
       pch=21, bg="olivedrab")
```

Ahora añadimos colores (véase la Fig. 5.17) para codificar por colores las distintas especies, porque sin diferenciación de subgrupos el diagrama de dispersión es de poca utilidad y no permite el reconocimiento de patrones.

```
# with color
pairs(iris[1:4], main="Iris data set (from Edgar Anderson)", pch=21,
      g=c("violetred2", "greenyellow", "blue")[unclass(iris$Species)])
```

Esto parece más amigable y permite un enfoque intuitivo. En la Figura 5.17 se observa que hay relativamente pocos solapamientos en casi todas las combinaciones de parcelas; por lo demás, es fácil distinguir tres grupos que corresponden exactamente a las tres especies. Sólo el gráfico de `Longitud.Sepal` frente a `Anchura.Sepal` muestra cierta ambigüedad en mayor medida. Por lo demás, los solapamientos se limitan a menos de un puñado de portadores de rasgos y éstos se limitan sólo a dos de las tres especies. La tercera especie está bastante aislada en todas las parcelas y, por tanto, está claramente delimitada. Hasta aquí la identificación estructural – la parte puramente AED es baja por el momento. Ahora sigue la clasificación con el análisis discriminante lineal `lda()` del paquete `MASS` de R. Más adelante, volveremos a AED cuando se trate de evaluar la calidad de la clasificación. calidad de la clasificación.

En primer lugar, se crean dos conjuntos de datos a partir del conjunto de datos del iris. El primero se utiliza para crear el modelo y el segundo se utiliza para la comprobación del modelo. La asignación es aleatoria y con `set.seed()` se puede repetir esta aleatoriedad si es necesario y reproducir exactamente la muestra. `set.seed()` permite la reproducción exacta de eventos "aleatorios", que no es nada más que un algoritmo que genera eventos distribuidos equitativamente a partir de un punto de partida libremente seleccionable. Por supuesto, esto no tiene nada que ver con el azar en el sentido de independencia de las condiciones causa-efecto o con el caos (que pueden ser diferentes lecturas de la aleatoriedad). En este sentido el concepto de azar debe descartarse por engañoso, ya que en el fondo se trata del resultado algorítmicamente determinado de las relaciones causa-efecto en función de las condiciones de entrada.

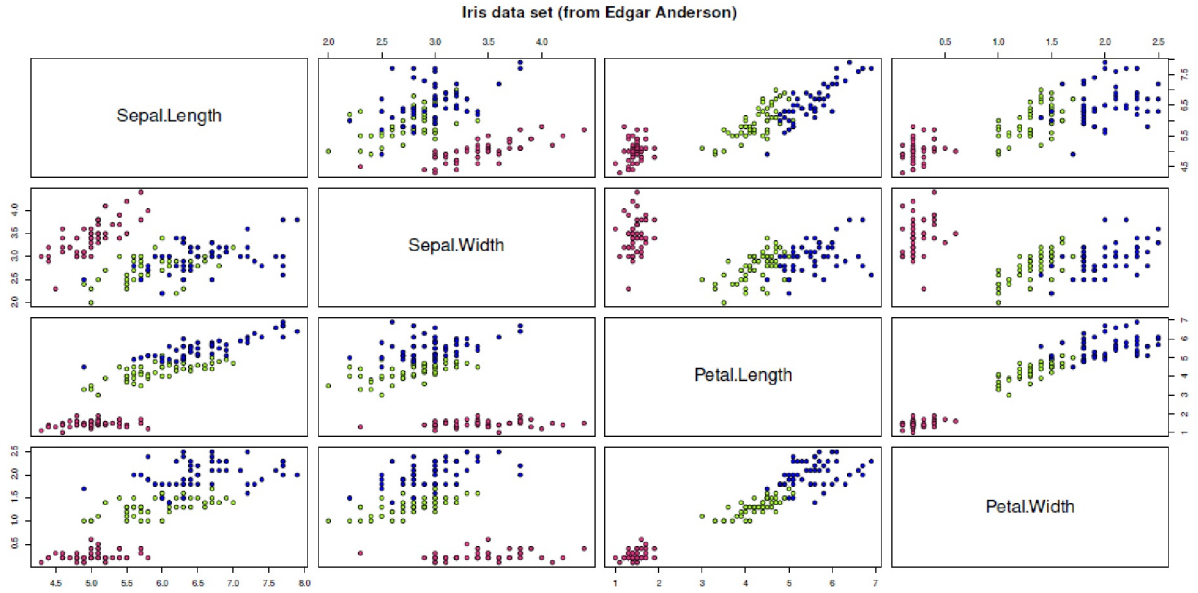


Figura 5.17. *Datos del iris (matriz de dispersión, especies coloreadas).*

```
# lda
# create train vs. test data subset from iris
seed <- 5642
set.seed(seed)
train.ids <- sample(di[1], di[1]/2)
train.iris <- iris[train.ids,]
test.iris <- iris[-train.ids,]
table(train.iris$Species)
table(test.iris$Species)
```

El LDA puede aplicarse a todo el conjunto de datos y al conjunto de datos de entrenamiento. A continuación se realiza la predicción en el conjunto de datos de prueba. A cada una de las especies se le asigna una probabilidad de entrada de $1=3$. Esto tiene sentido, ya que no se dispone de ninguna otra información y el número de datos está equilibrado (véase más arriba).

```
# prior = 1/3 for each species
# all
lda.iris <- lda(Species ~ ., iris, prior=rep(1/3,3))
# train data set
trainlda.iris <- lda(Species ~ ., train.iris, prior=rep(1/3,3))
# predict train data set
pred.iris <- predict(object=trainlda.iris, newdata=test.iris)
```

Ahora es el momento de mostrar los modelos ...

```
# check
lda.iris
trainlda.iris
str(pred.iris)
```

... y las clasificaciones erróneas para todos los datos, los datos de entrenamiento y los datos de prueba. Estos últimos son los relevantes.

```
> # all
> table(REAL=(actual <- iris$Species),
            + CLASSIFIED=(model.response(model.frame(lda.iris))))
CLASSIFIED
```

```
REAL       setosa versicolor virginica
setosa         50         0         0
versicolor      0        50         0
virginica       0         0        50
> # train data set
> table(REAL=(actual <- train.iris$Species),
+ CLASSIFIED=(model.response(model.frame(trainlda.iris))))
CLASSIFIED
REAL       setosa versicolor virginica
setosa         24         0         0
versicolor      0        26         0
virginica       0         0        25
> # predict for test data set - that is relevant
> pred.tab <- table(REAL=(actual <- test.iris$Species),
           + CLASSIFIED=(pred.iris$class))
> pred.tab
CLASSIFIED
REAL       setosa versicolor virginica
setosa         26         0         0
versicolor      0        22         2
virginica       0         4        21
```

La proporción de clasificaciones correctas e incorrectas se suma sencillamente a lo largo de la diagonal:

```
> # correct classified
> cclass <- sum(diag(pred.tab))/sum(pred.tab)
> # incorrect classified
> wclass <- 1- cclass
> cclass
[1] 0.92
> wclass
[1] 0.08
```

Un porcentaje de asignaciones correctas del 97.33% y de clasificaciones erróneas del 2.6% parece bastante respetable. Más concretamente, para la especie Virginica se clasifica un espécimen como Versicolor y para Versicolor un espécimen como Virginica. La especie setosa ha sido clasificada de forma completamente correcta. Recordatorio – *esta clasificación* se basa en una generación aleatoria de conjuntos de datos de entrenamiento y de prueba y depende del valor pasado a `set.seed()`. Los lectores interesados pueden cambiar el valor de `set.seed()` y repetir el análisis – un poco diferentes resultarán otros valores. Por lo tanto, para probar las tolerancias o la robustez del análisis, sería necesario repetir este proceso muy a menudo y generar muchos conjuntos de datos de entrenamiento o de prueba y anotar las clasificaciones correctas e incorrectas en cada caso. Esto conduce una estimación robusta de la calidad de la clasificación. Los valores dados aquí están sujetos a las fluctuaciones de una muestra aleatoria.

Para saber con mayor precisión qué especies se han clasificado correcta e incorrectamente, creamos una tabla comparativa.

```
# compare -- create tables
pred.iris.tab <- data.frame(pred.iris$x, test=test.iris$Species,
           pred=pred.iris$class)
comp.ids.wc <- with(pred.iris.tab, which(test != pred))
comp <- as.character(pred.iris.tab$test)
replacetext <- do.call(paste,
c(pred.iris.tab[comp.ids.wc,c("test","pred")],
           sep=":"))
comp[comp.ids.wc] <- replacetext
pred.iris.tab <- data.frame(pred.iris.tab, comp)
pred.iris.tab
pred.iris.tab$compTF <- TRUE
pred.iris.tab$compTF[comp.ids.wc] <- FALSE
head(pred.iris.tab)
tail(pred.iris.tab)
pred.iris.tab[comp.ids.wc,]
```

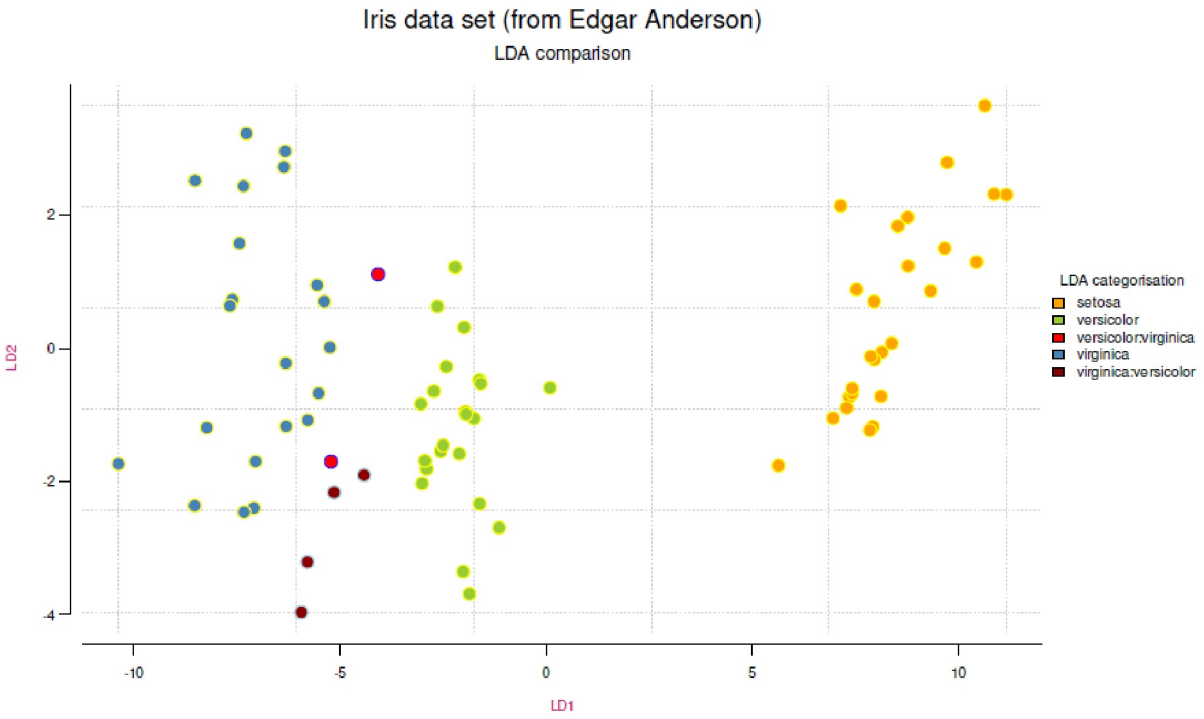


Figura 5.18. *Datos del iris (calidad de la clasificación)*

La calidad de la clasificación puede examinarse gráficamente mediante LDA. Los objetos LDA pueden trazarse con `pairs()` . Sin embargo, como aquí sólo se dispone de dos dimensiones, basta con una simple llamada a `plot()` (véase Fig. 5.18).

```
comp.fac <- as.factor(pred.iris.tab$comp)
# plot alone is sufficient
levels(comp.fac)
bgcolo <- c("orange", "yellowgreen", "red",
       "steelblue","darkred")[comp.fac]
       bocolo <- c("yellow", "yellow", "blue", "yellow",
       "lightblue")[comp.fac]
xlim <- range(pred.iris$x[,"LD1"])
ylim <- range(pred.iris$x[,"LD2"])
par(mar=c(5,5,4,9), oma=c(2,1,1,1), "cex.axis"=0.8)
plot(pred.iris$x, pre.panel=grid(), col.lab="violetred3",
       cex.lab=0.85, pch=21, cex=1.7, cex.lab=0.8, main="",
       cex.axis=0.8, col=bocolo, bg=bgcolo, axes=F,bty="n")
axis(side = 1, pretty(xlim), tck =- .02, labels=NA, line=.6)
axis(side = 2, pretty(ylim), tck = -.02, labels=NA, line=.6)
axis(side = 1, lwd = 0, line = .4)
axis(side = 2, lwd = 0, line = .4, las = 1)
mtext(TITLE, 3, line=2.5, cex=1.5)
mtext(SUB, 3, line=1, cex=1.1)
legend(12,1.2, as.vector(levels(comp.fac)),
       horiz=FALSE, xpd=TRUE,
       fill=c("orange", "yellowgreen", "red",
              "steelblue", "darkred"),
       title="LDA categorisation", cex=0.85,
       bty="n", title.col="black")
```

El conjunto tiene aún mejor aspecto con los nombres de las especies directamente en el gráfico (véase la Fig. 5.19).

```
# text plotten
fac <- 1.2
xlim <- range(pred.iris$x[,"LD1"])*fac
ylim <- range(pred.iris$x[,"LD2"])*fac
textplots <- pred.iris.tab$comp
par(mar=c(5,5,4,9), oma=c(2,1,1,1), "cex.axis"=0.8)
plot(0,0, xlim=xlim, ylim=ylim, pre.panel=grid(),
           pch=21, cex=1.7, cex.lab=0.8, cex.axis=0.8, col="white",
           xlab="", ylab="", bg="white", axes=F, main="", bty="n")
text(pred.iris$x, labels=pred.iris.tab[,"comp"], col=bgcolo)
axis(side = 1, pretty(xlim), tck =- .02, labels=NA, line=.6)
axis(side = 2, pretty(ylim), tck = -.02, labels=NA, line=.6)
axis(side = 1, lwd = 0, line = .4)
axis(side = 2, lwd = 0, line = .4, las = 1)
mtext(TITLE, 3, line=2.5, cex=1.5)
mtext(SUB, 3, line=1, cex=1.1)
title(xlab="LD1",ylab="LD2", col.lab="violetred3", cex.lab=0.85)
legend(15,2.4, as.vector(levels(comp.fac)), horiz=FALSE,
           fill=c("orange","yellowgreen","red","steelblue","darkred"),
           title="LDA categorisation", cex=0.85, xpd=TRUE,
           bty="n", title.col="black")
```

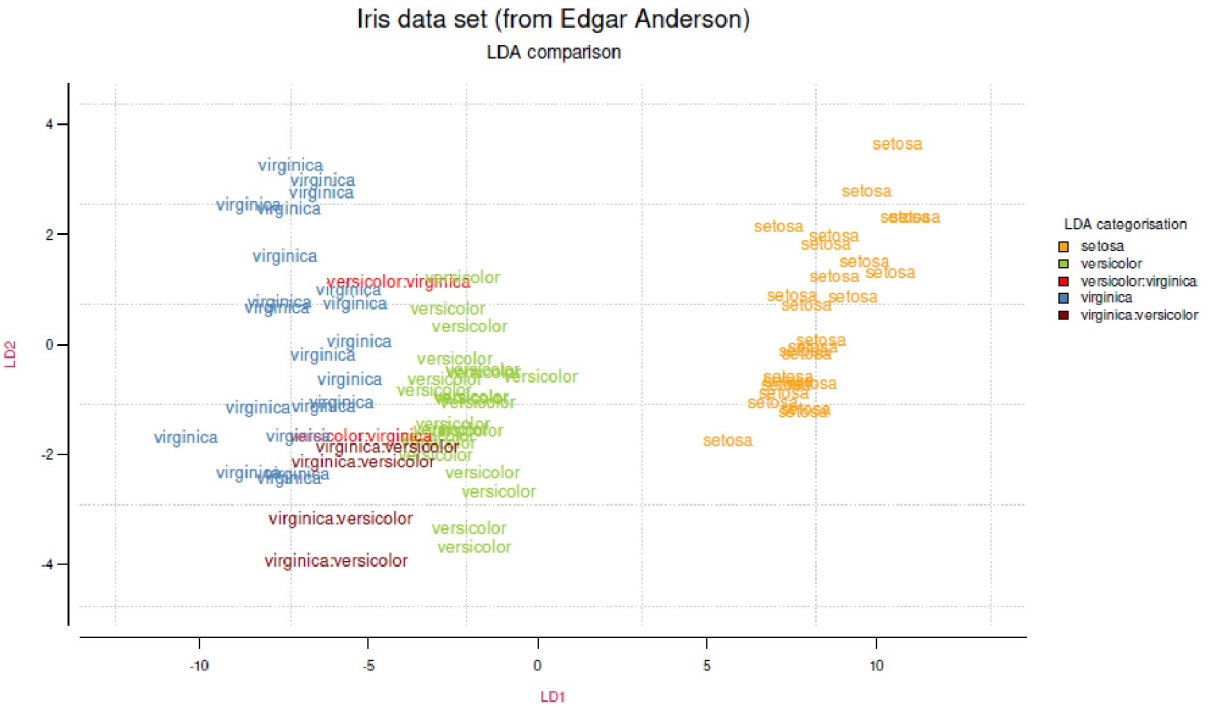


Figura 5.19. *Datos del iris (grado de clasificación, con el nombre de la especie)*

El texto y las marcas de color muestran qué especies se han clasificado correctamente y cuáles no. Lo que aún podría hacerse ahora sería revisar los nombres - por ejemplo, abreviar con `abbreviate()` o el uso de abreviaturas propias o la combinación de puntos y texto. Las abreviaturas podrían tener este aspecto:

```
> # abbreviate names
> abbreviate(as.vector(levels(comp.fac)))
setosa versicolor versicolor:virginica
"sets"     "vrsc"     "vrs:"
virginica virginica:versicolor
"vrgn"         "vrg:"
```

Como alternativo se podría comparar el análisis discriminante lineal con el análisis jerárquico de clusteres y optimizarlo como se muestra en la Figura 5.20. La norma en R – distancias euclídeas y el método de aglomeración completa – se utilizan para simplificar. El ejemplo muestra enfáticamente lo que debería explorarse ahora a lo largo de las Figuras 5.18, 5.19 y 5.20 respectivamente, cuál de los algoritmos de agrupación jerárquica por aglomeración es adecuado para estos datos y, sobre todo, por qué y cómo se relaciona con el análisis discriminante lineal para poder explicar las diferencias en los resultados.

```
# hclust
plot(hclust(dist(iris[,1:4])),
           labels=abbreviate(iris$Species,2), col="darkred")
```

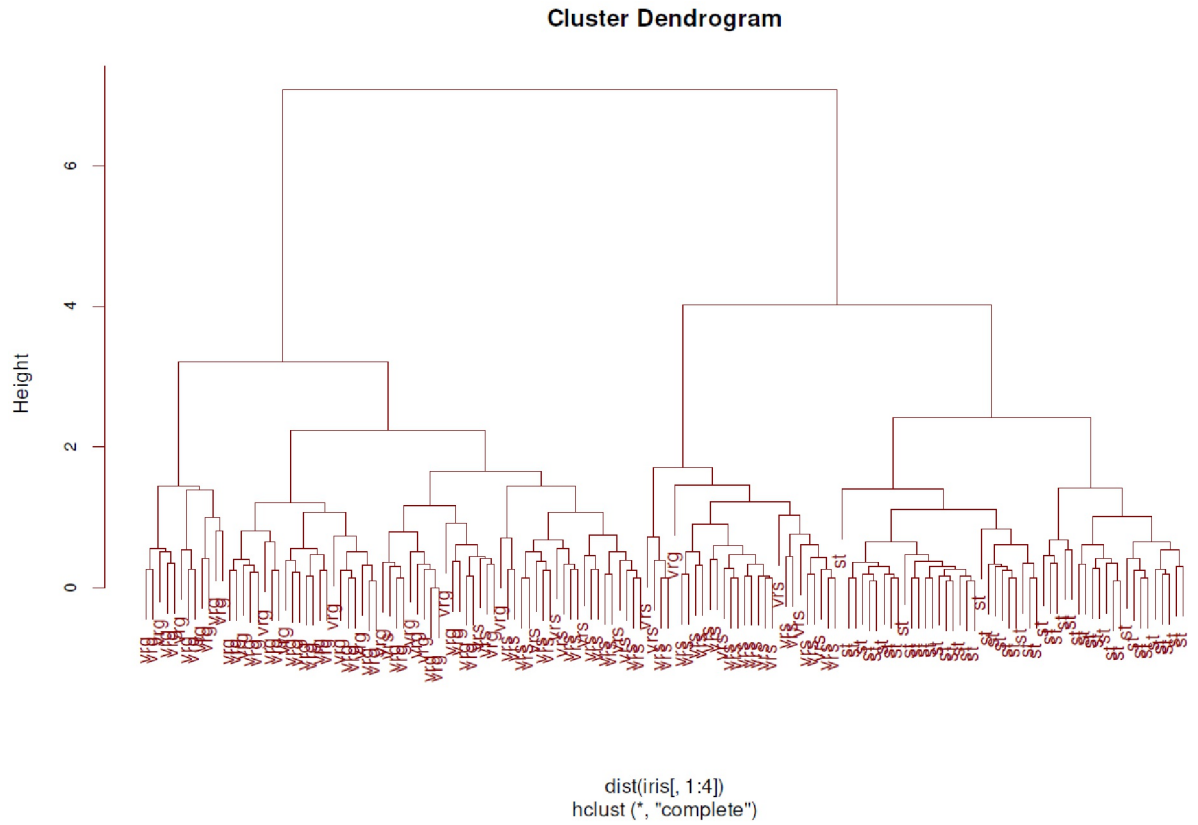


Figura 5.20. *Datos del iris (análisis jerárquico de clusteres)*

El mensaje de este ejemplo era que el AED -realizado mediante análisis gráfico- es muy bueno para la comprobación de modelos, y no sólo para modelos lineales y diagnósticos residuales. El AED puede tanto dar una impresión estructural de antemano como abrir variantes a posteriori a la hora de comprender los resultados o planificar los pasos siguientes. Además, los algoritmos de clasificación pueden ser una parte excelente del AED, si no se basan en significados y decisiones de prueba, sino en los resultados de las respectivas asignaciones y agrupaciones.

5.5.4 Vivir y morir a bordo del Titanic

El siguiente AED desarrolla el enfoque basado especialmente en consideraciones cualitativas para utilizar la información contextual. En el contexto bayesiano, esto podría utilizarse para convertir la información prioritaria en una prior matemática, ya que faltan datos comparativos de anteriores naufragios comparables.

5.5.4.1 Contexto

El hundimiento del barco Titanic el 15 de abril de 1912, que se anunciaba como insumergible, directamente en su primer viaje a América – e increíblemente por una colisión presumiblemente evitable con un iceberg – ha provocado que un gran número de personas se ocupen de este naufragio y sigan haciéndolo. Esto va mucho más allá de la conocida película de James Cameron y la búsqueda de objetos en el fondo del mar. Hay muchas páginas web con más información sobre el barco, la tripulación y los pasajeros, fotos originales o dibujos técnicos del barco y una reconstrucción de los procesos en el transcurso de la colisión y el hundimiento del barco. Incluso hay películas de animación, para que el hundimiento de la parte delantera del barco llegó al fondo relativamente bien conservada y la parte trasera fue desmantelada y destrozada casi por completo y luego se hundió hasta el fondo en dos grandes trozos (National Geographic, 2012; Gerber, 2012).

Hay suficientes datos – tantos, por ejemplo, sobre la tripulación y los pasajeros – que incluso hay competiciones para ver quién puede crear el mejor modelo predictivo para predecir con la mayor exactitud posible quién sobrevivió y quién no, basándose en las características disponibles de las listas de personas. Los modelos están escritos en su mayoría en R o Python y utilizan, entre otras cosas, partición recursiva y árboles de regresión (Breiman, Friedman, Olshen & Stone, 1993, paquete de R `rpart`) y otros algoritmos de clasificación, que no tratamos en detalle aquí.

5.5.4.2 Cuestionamiento

Dejamos de lado la cuestión de los sucesos reales y las tragedias personales en este análisis. Lo que nos interesa son las listas de tripulantes y pasajeros. Nos acercarnos a este conjunto de datos con el menor conocimiento posible – pero con razón – y consideramos qué escollos existen al trabajar con estos datos. Lo primero es – buscamos una pregunta para aplicar el AED en esta línea. Además de la pregunta muy general de "¿Quién sobrevivió?" podemos mirar más específicamente a:

1. ¿Se aplicó en el Titanic el lema "las mujeres y los niños primero" a la hora de salvar a la gente?
2. ¿Quién le hubiera gustado ser a usted a la hora de sobrevivir y quién no?
3. ¿En qué medida cuidaba el empresario Titanic de sus empleados?
4. ¿Qué patrones se pueden reconstruir a partir de la información contextual y de los datos sobre los pasajeros y la tripulación, respectivamente, para poder demostrar empíricamente la supervivencia y la muerte en el Titanic de forma comprensible y fundamentada?

Trabajaremos en todas las preguntas y veremos en qué medida nos ayudan los datos. Sin embargo, no analizaremos y responderemos a todas las preguntas hasta el más mínimo detalle ni trabajaremos de forma confirmatoria sino sólo en la medida en que toda la información y las aproximaciones a los datos estén disponibles a través del AED, para que los lectores puedan profundizar en el tema según sus propios intereses.

En primer lugar, reconstruimos muy brevemente y a grandes rasgos el contexto histórico de principios del siglo XX y las condiciones del Titanic basándonos en consideraciones generales. En principio, estas suposiciones pueden comprobarse al azar mediante datos empíricos o búsquedas en Internet, lo que sin duda requiere cierto esfuerzo. Nuestras explicaciones son breves y, por tanto, probablemente erróneas, pero deberían inspirarle para construir sus propias cadenas de argumentación con el fin de situar cualquier información contextual en un marco razonable. Es importante destacar que nos las arreglamos durante mucho tiempo sin mirar directamente ninguno de los datos anunciados, es decir, las listas de tripulantes y pasajeros. Más bien, *reconstruimos el caso*. Desgraciadamente, este importantísimo proceso de reconstrucción del caso se discute demasiado poco en los artículos y se descuida y no se describe en los libros de texto (de estadística), cómo proceder. Sin embargo, el análisis del caso es esencial antes de poder cribar los datos porque es el punto a partir del cual planificamos los pasos e investigaciones posteriores. Para el análisis cualitativo de casos nos guiamos por Oevermann (2000), Hildenbrand (1999) y Wernet (2000).

5.5.4.3 Reconstrucción del contexto histórico

En el periodo previo a la Primera Guerra Mundial, la primera década del siglo XX se caracteriza por la emigración (por ejemplo, a Occidente/EE.UU., tanto de familias como de individuos), la industrialización y las visiones nacionalistas o imperialistas-reales-elitistas en Europa. Al mismo tiempo, se aproxima la Revolución Rusa en el Este y la Primera Guerra Mundial a escala mundial o en Europa, con el consiguiente colapso de reinos (por ejemplo, Austria-Hungría, Alemania, etc.). El contraste entre ricos y pobres, o más bien entre los emergentes pobres industriales y los ricos industriales en el contexto de las ciudades, sugiere un reflejo de estas condiciones en el Titanic. Así, se puede investigar cuánta diferencia había entre las tres clases de tarifas y, además, cuán grande era la diferencia entre las suites caras y los camarotes normales, una diferencia enorme. Paralelamente, el Titanic fue concebido como un vínculo entre el viejo mundo europeo y el nuevo mundo americano y el mundo como tal - antes de la época de los aviones – significativamente empezando por el paso más rápido a América a través del Atlántico Norte. Así pues, puede considerarse que el Titanic representa un estilo de vida elitista y, al mismo tiempo, el progreso tecnológico y la globalización. Además de las tragedias personales, el hundimiento es precisamente una pieza de la vida público-social, ya que se discutió mucho a través de los periódicos y la prensa escrita sobre por qué se hundió el Titanic, quién tuvo la culpa y qué comportamiento pudieron observar a bordo durante la catástrofe las distintas personas y representantes del barco. Como ocurre con todos los acontecimientos históricos ampliamente debatidos, no existe una opinión uniforme al respecto: cada superviviente percibió este acontecimiento de forma diferente. Hoy en día, todos estos temas están mucho más rápidamente presentes a través de Internet que en la época de los periódicos, sin televisión y, como mucho, con una radio. Sin embargo, las reacciones de la época están a la altura de lo que hoy se denomina "tormenta de mierda", y el Senado de Estados Unidos realizó una investigación oficial (19.04.-25.05. 1912, Titanic Inquiry Project, 1998), que, sin embargo, no llegó a nada y se basó en relaciones de poder. El hecho de que el propio capitán – Edward John Smith (1850-1912) – se hundiera y muriera con el Titanic fue probablemente "útil" para este proceso. El público, por supuesto, discutió el acontecimiento en detalle.

El Titanic se consideraba insumergible y éste era su primer viaje. Los costes de producción costaron la asombrosa cifra de 1,5 millones de libras, lo que, según un post de Turk (2017), se traduce en conversión de 1.660 millones de dólares en la actualidad. El Titanic se construyó junto con sus buques gemelos, el más pequeño RMS Olympic de 1911 y el HMS Britannic completado en 1914 (que, por cierto, también se hundió como buque hospital en 1916, presumiblemente a causa de una mina o un torpedo). El Titanic fue también el mayor buque de pasajeros de la época. Tenía 270 m de eslora y hasta la chimenea 53 m de altura y desplazaba 54 000 t con un calado de ~10,54 m. Vista de estas dimensiones, había ciertamente una cierta arrogancia hacia el mar y las fuerzas de la naturaleza – característica de la idea aún imperante de "superioridad" humano-técnica sobre la vida misma y sobre las fuerzas de la naturaleza. Este pensamiento elitista fundamental tuvo – según nuestra tesis – graves consecuencias para la privilegiada de los camarotes, el acceso a los recursos (por ejemplo, los botes salvavidas), la información (por ejemplo, que se les avisara a tiempo de que el barco se hundía, de que la situación era grave, etc.). Los trabajadores ordinarios eran una mercancía de masas, sobre todo en el transcurso de la Revolución Industrial, mientras que las mujeres y los niños siempre han tenido un papel privilegiado en las catástrofes. En la práctica, sin embargo, suele resultar que afirmaciones tan rotundas rara vez son generalizables. ¿Qué significaba esto para las perspectivas de sobrevivir en el Titanic? ¿Significó que todas las mujeres y los niños fueron los primeros (en los botes salvavidas) y que las pérdidas fueron mayores en las clases más bajas y entre la tripulación, si lo operacionalizamos según un índice económico? Los lectores perdonarán tal terminología – el uso de consideraciones económicas no niega las muertes de personas y las tragedias y destinos asociados a ellas, sino que pero puede ayudar a aclarar algunas cosas.

5.5.4.4 Datos sobre el Titanic

5.5.4.4.1 Capacidad y equipamiento . Había más de 2202 personas a bordo cuando se hundió el Titanic. Esto se desglosa en ≈1317 pasajeros y ≈ 885 miembros de la tripulación – dependiendo de la fuente (incluyendo Wikipedia, 2019c; Encyclopedia Titanica, 1996). Una parte considerable de la tripulación (≈325 personas)

trabajaba en la zona de máquinas, es decir, en las profundidades del casco; entonces todo era trabajo manual, no había ordenadores ni unidades de control eléctrico. Una parte más numerosa (≈500 personas) trabajaba en la zona del hotel, principalmente como camareros o azafatas. Sin embargo, el Titanic no estaba completo cuando se hundió. En principio, podría haber transportado a 3.327 personas. Sólo disponía de 20 botes salvavidas, con capacidad para 1178 personas. Los botes estaban situados en la cubierta superior, cerca de la primera y segunda clase. La falta de botes salvavidas era legal en aquella época, porque el número de botes salvavidas se basaba en el tamaño del barco y no en el número de pasajeros y tripulación. Hoy, afortunadamente, tal cosa ya no estaría permitida. Cabe suponer que en vista de sólo 1178/3327*100% = 35.41% disponibles botes salvavidas (a ocupación máxima) o 1178/2202*100% = 53.5% disponibles botes salvavidas (= ocupación en el momento del hundimiento), la probabilidad global de supervivencia se redujo en un 46.5%, es decir, ¡se redujo a la mitad! Y esto sólo era cierto si se botaban todos los botes y se informaba a tiempo a todas las personas y, en consecuencia, se llenaban todos los botes salvavidas disponibles. Sin embargo, sólo se utilizaron 18 de los 20 botes salvavidas. La rotura del casco redujo la probabilidad de supervivencia en otro tanto por ciento. El agua helada a -2 °Celsius (ligeramente por encima del punto de congelación del agua de mar) – un salto al agua con chaleco salvavidas y sin sitio en un bote salvavidas – sólo permitía unos minutos de supervivencia (15-30 min. como máximo). El siguiente barco estaba tan lejos que no llegó hasta las 4 de la madrugada. Eso fue más de cuatro horas después de la colisión. Independientemente de esto, la supervivencia dependía de otros factores que podemos examinar provisionalmente utilizando las características de las personas disponibles. Las características sociodemográficas influyen en el nivel de la masa, pero no necesariamente en la suerte individual. Así pues, las declaraciones se refieren a todos en paralelo (pasajeros y tripulación) y los casos individuales pueden ser excepciones, pero no cumplen los requisitos de una norma. Para el análisis se utilizarán las listas de tripulantes y pasajeros.

5.5.4.4.2 Proceso de hundimiento. El barco fue embestido por el iceberg en la parte delantera de estribor (= a la derecha en el sentido de la marcha) y se abrió por el costado, de modo que la proa se llenó lentamente de agua y el barco se hundió hacia delante en el mar. Antes de hundirse del todo, se partió más bien por la zona central, lo que hizo que la proa y la parte delantera se hundieran como una pieza relativamente entera a gran velocidad hasta el fondo del mar, y la parte trasera – al estar abierta, desprotegida y completamente indefensa sin proa, quedó más o menos desgarrada y desmantelada. Estas piezas se desgarraron aún más con el hundimiento y se esparcieron por una gran superficie del lecho marino. Es decir, los botes salvavidas que estaban en la zona trasera en el corrían peligro de ser alcanzados por los restos del naufragio o por la rotura del casco. Las personas que nadaban se encontraban así indefensas y en un peligro aún mayor. Las personas que se encontraban en los botes salvavidas aún podían morir en esta zona debido a la rotura del casco y al maremoto resultante, a pesar de ser supuestamente rescatadas en el bote, ya fuera debido a las heridas provocadas por los restos del naufragio, por la entrada de agua en los botes salvavidas o simplemente debido al frío o al agotamiento. Esto último se aplicaba igualmente a todas las personas que no se hundieron directamente con el Titanic.

5.5.4.4.3 Hora del día. El iceberg golpeó al Titanic por la noche alrededor de las 23:40 y se hundió después de Δt = 2:40 horas a las 02:20 de la mañana (hora del barco), lo que afectó negativamente al estado de alerta general de todos los pasajeros (incluyendo oscuridad, posible pérdida de orientación, pánico, tener sueño por la noche, bajas temperaturas exteriores, etc.). No está claro y presumiblemente no se dispone de datos, quién podría haber tenido ventaja en estas circunstancias – quizás personas que habitualmente estaban despiertas por la noche o que trabajaban de noche.

5.5.4.4.4 Disposición espacial. Tradicionalmente, los camarotes caros se encontraban en el piso superior, los baratos en el piso inferior y los camarotes de la tripulación presumiblemente más abajo en el casco. Esto podría verificarse con planos de distribución del Titanic, cosa que no hemos hecho aquí. Sin embargo, la proximidad espacial puede haber jugado un papel, además de la situación de quién estaba dónde y cuándo exactamente el Titanic chocó con el iceberg y comenzó a hundirse. Así que podría haber diferencias en la tripulación, si un azafato / azafata estaba trabajando en primera clase o un obrero en lo profundo de la sala de máquinas ... Esto sería difícil de reconstruir.

5.5.4.4.5 Información. Suele transmitirse en primer lugar a los privilegiados. Los trabajadores, por su parte, suelen recibir mucha información "de tapadillo" mientras trabajan. Ahora se plantea el problema del "correo silencioso": una persona cuenta una historia a otra, que hace lo mismo, y así sucesivamente. Con cada relato posterior, la historia se modifica ligeramente de forma más o menos deliberada. Después de que la misma historia se haya contado varias veces, a menudo sólo tiene una conexión rudimentaria con los orígenes y la información "real", e incluso entonces no está claro si la primera versión de una historia era exacta en absoluto. Este puede ser un motivo para la propagación de la desinformación; y dejamos fuera deliberadamente la desinformación intencionada. No obstante, podemos suponer con cautela, a partir del contexto, que primero se prestó atención a los pasajeros de clase superior antes de pensar en la tripulación. Esto podría verificarse cualitativamente mediante registros históricos de supervivientes y acontecimientos reconstruibles durante el hundimiento.

5.5.4.4.6 Comportamiento en las crisis. Otro factor de influencia relevante es que las personas en crisis tienden a menudo a pensar aún más en sí mismas o a no pensar en absoluto y sólo actúan y su percepción se estrecha drásticamente, de modo que incluso las cosas sencillas no son (o dejan de ser) posibles. Así que cabe suponer que no se utilizaron todos los botes salvavidas y que no estaban totalmente tripulados. Ambas cosas requerirían una logística perfecta y una actitud sistemática de "mantener la calma", en el sentido de que, dada la falta total de botes, deberían haberse llenado algo más, pero no demasiado, para salvar a más personas, pero sin poner en peligro los botes por hacinamiento. Del mismo modo, puede haber negación y la gravedad de la situación sólo se tiene en cuenta parcialmente, de modo que no se informa con la suficiente rapidez. De este modo, la crisis como factor en sí mismo vuelve a reducir considerablemente la probabilidad de supervivencia. Todos los puntos tratados son precisamente los procesos que conducen al conocimiento previo (O'Hagan, Buck, Daneshkhah, Eiser, Garthwaite, Jenkinson, Oakley, & Rakow, 2006) y dan lugar a una distribución a priori, no se trataría de AED, sino de Bayes y de la transformación de información cualitativa en una distribución a priori (véase el capítulo 6.12).

5.5.4.4.7 Recursos. ¿Qué pasa ahora con los recursos? Sería bueno contar con una tripulación experimentada en crisis, pero probablemente no era el caso. El experimentado capitán Edward John Smith (1850-1912) dirigió el barco, pero ¿se tomó la situación lo suficientemente en serio y actuó de acuerdo con las necesidades al margen de su propio ego? La cooperación entre personas suele considerarse una ventaja evolutiva (Nowak, 2006; West, Ashleigh & Gardner, 2007; Boyd & Richerson, 2009). Estar solo lo hace mucho más difícil en las crisis. En casos individuales, la falta de cooperación puede tener un efecto positivo, por ejemplo, si uno tiene que desaparecer, lo que es más difícil para un grupo grande, o es mejor no coordinarse para ser más flexible. Estar en un grupo demasiado grande y dejar de ser flexible puede, a su vez, anular la ventaja de supervivencia que realmente se deriva de la cooperación, según el lema "una cadena es tan fuerte como su eslabón más débil". Un grupo que no sea demasiado grande debería garantizar una ventaja de supervivencia, así que no viajes solo o con demasiada gente. Las familias más pequeñas o madre e hijo(s) podrían obtener aquí una pequeña ventaja sobre otros grupos debido a las ventajas de la empatía y los valores sociales. Sin embargo, habría que comprobar la realidad subyacente y probablemente no pueda considerarse independiente de la clase social. Son posibles otras consideraciones, pero las dejamos a los lectores.

5.5.4.5 Supuestos y tesis

Resumamos ahora las observaciones anteriores sobre los supuestos y las tesis de la supervivencia relativa:

5.5.4.5.1 Estatus social. Los pasajeros de las clases más altas tenían más posibilidades de sobrevivir que los de las clases más bajas. Dependiendo de la personalidad de las personas de las clases más altas, podía ocurrir que en algunos casos los botes salvavidas estuvieran bloqueados por ellos y no estuvieran lo suficientemente llenos. Sin embargo, debe tratarse de casos aislados y poco sistemáticos, ya que probablemente faltó cierta sistematicidad en el pánico. Más sistemática, sin embargo, puede haber sido la cuestión de la difusión de la información. También en este caso sospechamos que los pasajeros de las clases más altas disfrutaron de una ventaja de supervivencia relativamente mayor.

5.5.4.5.2 Tripulación. Los miembros de la tripulación tuvieron menos posibilidades de sobrevivir que los pasajeros de clases superiores e incluso puede que estuvieran en peor situación que los pasajeros de tercera clase. En primer lugar, muchos se encontraban en las salas de máquinas muy por debajo de las cubiertas, al igual que probablemente sus camarotes, luego muchos estaban trabajando, etc. Por el contrario, eran camareros y, por tanto, estaban en todas partes del barco. Esto debería equilibrarse aproximadamente dado el gran número de tripulantes en la zona del hotel. En principio, esto podría haber permitido a muchos tener una oportunidad de rescate. Además de la información de que realmente tenían que abandonar el barco, también eran necesarios tanto la forma de subir como la de salir del lugar de trabajo, así como el libre acceso a los botes salvavidas. No hay que subestimar estos últimos argumentos. Podría ser que en el caso de una actitud honorable hacia el trabajo y/o la identificación con el Titanic, algunas personas pensaran que aún podían detener la catástrofe – por ejemplo, bombeando agua – y por ello cumplieron con su deber y murieron. Sin embargo, es posible que se trate de casos aislados. Un análisis histórico podría aclararlo.

5.5.4.5.3 Proximidad espacial. Cuanto más lejos (espacialmente) de los botes salvavidas – y esto es lo único que importa en un naufragio, ya que en aguas heladas se sobrevive muy poco tiempo a pesar del chaleco salvavidas – menores son las posibilidades de supervivencia. Si los botes salvavidas estaban en la zona de primera y segunda clase, es una clara ventaja estar cerca de estos botes. Es cuestionable el acceso de la tercera clase a la cubierta de la primera y segunda clase y si había barreras, lo que era habitual (miedo a la transmisión de enfermedades, separación de clases en términos de privacidad, etc.). Tampoco está claro el acceso de los distintos miembros de la tripulación a los botes salvavidas. Además, habría que analizar temporalmente quién fue informado, cuándo y con qué frecuencia sobre el estado de la situación y a qué ritmo se bajaron al agua los botes salvavidas en relación con la duración total del proceso de hundimiento.

5.5.4.5.4 Transmisión de información. Cuanto más "lejos" del mando del barco, menos fiable es la información que llega y señala el camino directo y más corto hacia los botes salvavidas e informa para abandonar realmente el barco de inmediato. Las únicas excepciones son las personas a las que se informa habitualmente, posiblemente personas de las clases superiores, así como mujeres y niños, pero que entonces deben ser conocidos, lo que probablemente era más el caso en las clases superiores.

5.5.4.5.5 Obstáculos. No se utilizaron todos los botes salvavidas – las razones carecen de importancia aquí –, lo que redujo las posibilidades individuales de supervivencia. A esto se sumó la falta de botes salvavidas y muchos procesos perceptivos potencialmente distorsionadores (por ejemplo, negación, ignorar advertencias) por parte de la dirección del barco. Tras el hundimiento, también se plantea la cuestión de si un barco ocupado regresará y recogerá activamente a los nadadores. El miedo a ser "invadido" por la "masa de ahogados", es decir, a que el propio barco se llene demasiado y se hunda, puede tener un efecto inhibidor. Es probable que esto tenga un efecto negativo en el comportamiento de ayuda, incluso si objetivamente hubiera habido suficiente espacio en los botes. También es probable que la intención humana sea alejarse del lugar del accidente, lo que no habla en favor de regresar y buscar activamente supervivientes en el agua. Del mismo modo, puede existir un temor fundado a que el barco que aún se hunde o los restos del naufragio puedan dañar los propios botes salvavidas. Las excepciones son posibles y esperables; las desviaciones sistemáticas parecen poco probables.

Caso 5.1: Posibilidades de supervivencia en el Titanic

¿Fueron las posibilidades de supervivencia en el Titanic justas y se distribuyeron por igual para todos? La respuesta ya debe ser definitiva: ¡NO! Las posibilidades de supervivencia no estaban en absoluto distribuidas equitativamente y para muchos grupos ciertamente cualquier cosa menos justas. Entre ellos contamos a las clases bajas y a la tripulación.

5.5.4.5.6 Tamaño del grupo. Las personas que viajaban solas tenían menos ventaja para sobrevivir que las que lo hacían en familia o en grupo. Sin embargo, las familias o los grupos demasiado grandes (por ejemplo, de más de 4 ó 5 personas en adelante) tenían menos ventaja de supervivencia porque la coordinación y la logística se hacían demasiado pesadas. Así, en una crisis, se podía buscar a las personas no encontradas, con lo que no sólo se perdía un tiempo valioso, sino que se evitaba por completo el abandono. Los grupos pequeños tenían una mayor ventaja de supervivencia y los que viajaban solos tenían una ventaja de supervivencia ligeramente reducida en comparación con éstos.

5.5.4.5.7 Familias. Las mujeres, las madres y los niños pequeños tenían una mayor ventaja relativa de supervivencia con la misma clase o estatus social, así como la disposición de los camarotes y la distancia a los botes salvavidas, o con el mismo estatus de información. No está claro hasta qué punto la clase social tuvo un efecto, especialmente en las clases más bajas.

5.5.4.5.8 Asignación. Si la asignación de camarotes y su proximidad a los botes salvavidas está relacionada con el momento en que las personas subieron al barco, es decir, en qué puerto, entonces el embarque influye en la supervivencia posterior. Este sería el caso si los camarotes que estaban cerca de los botes salvavidas y más arriba se ocuparan primero, independientemente de la clase. Esto tendría que verificarse históricamente.

Todas las suposiciones anteriores pueden resumirse de forma condensada siguiendo los argumentos anteriores, y sin echar un solo vistazo a la lista de pasajeros y tripulantes hasta el momento. Sólo se utilizó la información generalmente conocida sobre el Titanic y un poco de pensamiento caso-reconstructivo. Es importante estar dispuesto a desprenderse de las propias suposiciones. Resumamos:

5.5.4.6 Conjuntos de datos

Hasta aquí las tesis: como siguiente paso necesitamos un conjunto de datos adecuado. Resulta que R tiene uno para el Titanic (Dawson, 1995). Veámoslo ():

```
?Titanic
Titanic
```

Como apunte - en R hay varios registros diferentes más para Titanic, como puede averiguarse:

```
help.search("titanic")
```

Por desgracia, el conjunto de datos `Titanic` de R ya está agregado como una tabla. No son datos brutos y faltan algunas variables importantes. Por lo tanto, sólo los datos de resumen sobre la clase, sexo, grupo de edad y supervivencia. Nos gustaría tener algo más. Pero antes, por curiosidad veamos gráficamente el conjunto de datos y dejemos que nos afecte:

```
# just have a look on the 'Titanic' dataset and forget this immediately!
mosaic(Titanic, shade=TRUE)
assoc(Titanic, shade=TRUE)
```

Y ahora practicamos algo nuevo - nos entrenamos en una habilidad que es inmensamente útil en ciencia: olvidamos completamente lo que acabamos de ver y desvanecemos conscientemente esta información y empezamos de nuevo desde cero y sin prejuicios. Las ilustraciones y sus explicaciones, que sólo aparecen impresas a continuación, se comentarán y revisarán más adelante. Mirar demasiado rápido los datos y su complejidad puede llevar a sacar conclusiones precipitadas y a comprometerse. Sin embargo, queremos encontrar patrones. Por lo tanto, empezamos modestamente y en pequeño con la información esencial. Basándonos en el análisis contextual anterior, ampliamos sucesivamente nuestro modelo y nuestros puntos de vista a lo largo de los datos empíricos. Al hacerlo, el AED siguiente prescindirá de la comprobación estadística de los valores característicos y, sin embargo, llegará a tesis bastante razonables. En nuestra opinión, un proceso de AED es, ante todo, un debate cualitativo, cuyas partes puramente cualitativas (por

ejemplo, la reconstrucción de la información cualitativa del caso) no suelen estar documentadas y, por lo tanto, rara vez se encuentran en la bibliografía, especialmente en los libros de texto de estadística. Los métodos de trabajo como la hermenéutica objetiva (véase el capítulo 11) lo hacen habitual y podemos aprender mucho de ellos. El análisis contextual es extremadamente útil, ya que estructura el proceso metodológico.

Sin embargo, lo que aprendemos del conjunto de datos del Titanic en R es que parece ser relativamente completo ya que enumera a 2201 personas.

```
> # everything there?
> sum(Titanic)
[1] 2201
> # for classes and crew
> apply(Titanic, c(1), sum)
1st 2nd 3rd Crew
325 285 706 885
> # classes
> sum(apply(Titanic, c(1), sum)[-4])
[1] 1316
```

5.5.4.6.1 Preparación del conjunto de datos. Para nuestro caso actual, utilizamos los datos disponibles públicamente sobre la tripulación del Titanic de Wikipedia (2019f) y los datos sobre los pasajeros de la Universidad de Vanderbilt, Departamento de Bioestadística, alojados por Frank E. Harrell (2002). Anteriormente estaban disponibles como dataset `titanic3` en el paquete R `Hmisc`. La entrada de Wikipedia también proporciona datos sobre los pasajeros. Sin embargo, allí falta información, por lo que se utiliza dicho otro conjunto de datos. Sería aún más preciso copiar los datos de la Encyclopedia Titanica (1996) y prepararlos en consecuencia.

En nuestro conjunto de datos para el análisis figuran ≈ 858 miembros de la tripulación y ≈1318 pasajeros, es decir, un total de ≈ 2176 personas. Esto significa que es muy probable que esté incompleto en lo que respecta a los miembros de la tripulación, ya que el conjunto de datos del Titanic en la R enumera 2201 personas. Por lo tanto, nuestro conjunto de datos solo contiene miembros de la tripulación conocidos por su nombre y que realmente estaban a bordo del Titanic (Wikipedia, 2019f; Encyclopedia Titanica, 1996). La preparación de datos a menudo requiere una cantidad considerable de trabajo con los datos antes de que algo esté siquiera disponible para su posterior análisis. Si se quiere ser muy preciso, es una buena idea cotejar diferentes fuentes de Internet entre sí. Las fuentes sobre el Titanic que se pueden encontrar en internet parecen ofrecer conjuntos de datos relativamente completos, pero en sentido estricto deben compararse por persona antes de agregar y utilizar las partes. Además, algunas fuentes se limitan a copiar otras más primarias. Por tanto, hay que evaluar la originalidad de las fuentes.

En el conjunto de datos del Titanic, integrar los datos significa dividir el apellido y varios nombres y añadir el título del saludo (por ejemplo, „Mr.“,“Mrs“,“Master“,“Miss“") primero en el contexto histórico – un "Master", por ejemplo, era un joven soltero – y luego, a partir de ahí, comparar qué combinaciones de apellidos, títulos y nombres se dan en los distintos conjuntos de datos y son, en definitiva, idénticas o no. R dispone de la función `merge()`, que integra las tablas por medio de columnas seleccionadas a través de las filas - y además puede emitir filas no integradas. Éstas deben examinarse manualmente, es decir, si la integración que falta se debe a errores tipográficos y se trata de la misma persona o si posiblemente se trata de dos personas diferentes con una gran similitud de nombres. Es útil convertir los nombres en minúsculas para anular las diferentes grafías. Una vez que haya integrado todos los nombres de modo que esté seguro de que la información coincide realmente, debe examinar la diversa información adicional (por ejemplo, identificación del bote salvavidas, embarque en el buque, tarifa, clase, sexo, edad, etc.) y comprobar si hay incoherencias. Así, un conjunto de datos puede tener valores ausentes (= NA) aquí y otro allá.

La integración pretende minimizar el número total de valores que faltan. Algunas fuentes de Internet no proporcionan toda la información disponible, sino sólo parte de ella, por lo que es necesario este trabajo manual y laborioso, pero a la larga merece la pena.

Tabla 5.2: *Comparación de las fuentes de datos (Titanic)*

Fuente	Datos	1st	2nd	3rd	Trip.	Σ
R	`Titanic`	325	285	706	885	2201
Frank Harrell (Univ. de Vanderbilt)	`titanic3.xls`	326	283	709		
WIKIPEDIA	copiado de pág. WEB				858	2176
	Diferencia	+1	-2	+3	-27	-25

Para ambos conjuntos de datos, los nombres y títulos se dividieron manualmente en un editor de texto con *buscar & sustituir* semiautomática y los nombres de las columnas se ajustaron con un programa de hoja de cálculo para poder trabajar con una única tabla de gran tamaño formada por los datos de pasajeros y tripulación. En principio, todos estos ajustes podrían realizarse con un código R en lugar de con un editor de texto. El procedimiento concreto depende de las preferencias y costumbres de cada uno. En el caso que nos ocupa, hemos optado por un editor de texto y un programa de hoja de cálculo. No se llevó a cabo una integración manual real entre diferentes conjuntos de datos, pero habría sido posible sin más, tal y como se ha descrito. La tabla 5.2 muestra la comparación en cuanto a la exhaustividad de nuestro conjunto de datos con el ya agregado conjunto de datos Titanic de la R. Las diferencias sólo son notables en el caso de los miembros de la tripulación. Estas podrían resolverse sin duda con un análisis de los datos de Wikipedia (2019c) y Encyclopedia Titanica (1996), respectivamente. Si tuviéramos que publicar un estudio dedicado a los datos del Titanic en una revista, sin duda habríamos asumido este esfuerzo de revisión manual sin dudarlo con el fin de aprovechar la mejor base de datos posible para todas las conclusiones. A efectos de demostración del proceso AED, esto no es necesario e ignoramos las discrepancias. Éstas representan algo menos del 1.15% de los datos y deberían ser lo suficientemente precisas como para inspirarse en las técnicas AED.

5.5.4.6.2 Conjunto de datos y variables existentes. El conjunto de datos puede leerse como .xlsx con el paquete R `xlsx` o `openxlsx` mediante `read.xlsx()`. También se puede seleccionar una tabla específica, aquí `combwithvanderbilt`. Primero hemos convertido la tabla del formato .xlsx a un formato basado en texto.

```
# import data
t.src <- read.table(file="Titanic_data.tab", header=TRUE, sep="\t")
dim(t.src)
str(t.src)
```

Alternativamente, leemos los datos de una tabla simple basada en texto con `read.table()`. Tenemos datos sobre n = 2176 personas y k = 25 variables, por lo que no todas las variables son relevantes. Algunas variables como la posición (= descripción del trabajo de los miembros de la tripulación) requieren cierto trabajo para ser tabuladas y, por lo tanto, no se siguieron teniendo en cuenta.

La tabla 5.3 explica las variables de interés.

Tabla 5.3: *Variables y valores (Titanic)*

Variable	Valor	Significancia	Tipo
sex	female, male	Sexo	factor
age	en años	Edad en el momento de accidente	numeric
sibsp	0, 1, 2, 3, 4, 5, 8	Número de hermanos o cónyuge a bordo (familia, hermano/cónyuge)	integer
parch	0, 1, 2, 3, 4, 5, 6, 9	Número de padres/hijos (familia, padre/madre/hijo)	integer
fare	en libras esterlinas	Tarifa	numeric
cabin	Cabinas ID	Cabinas ocupadas ID	factor
embarked	B, C, Q, S	Embarque en (B)elfast, (C)herbourg, (Q)ueenstown, (S)outhampton	factor
boat	Barco ID	Barco salvavidas ID	factor
body	Cuerpo ID	Cuerpo identificado	factor
pclass	Crew, First, Second, Third	Afiliación de la tripulación o del pasajero (y luego clase)	factor
survived	VERDADERO, FALSO	¿Sobrevivió el desastre?	logic

Primero, obtenemos las variables que nos interesan de la tabla original `t.src` y creamos nuevas a partir de ellas. Podríamos, por supuesto, trabajar con `attach(t.src)`, pero la experiencia demuestra que demasiados "attach" en paralelo llevan a confusión en cuanto a qué variables están actualmente activas y disponibles. Y así se tarda menos en teclear para acceder a las variables. Al mismo tiempo las clases de variables desfavorables se convierten (por ejemplo, a numéricas), de modo que cada variable esté disponible de tal forma que todos los análisis posteriores sean posibles.

```
# extract vars
sex <- t.src$sex
age <- as.numeric(t.src$age)
sibsp <- t.src$sibsp
parch <- t.src$parch
fare <- as.numeric(as.character(t.src$fare))
cabin <- t.src$cabin
embarked <- t.src$embarked
pclass <- t.src$pclass
survived <- t.src$survived
fullname <- as.character(t.src$fullname)
lastname <- as.character(t.src$lastname)
fulltitle <- as.character(t.src$title)
firstname <- as.character(t.src$firstname)
secondfirstname <- as.character(t.src$secondfirstname)
```

La mayoría de las variables se almacenan como factores. Edad (*age*) y tarifa (*fare*) las necesitamos numéricamente y las convertimos en consecuencia. Varias fuentes sugieren crear una nueva variable a partir de *sibsp* y *parch* en una nueva variable, *famsize*, que combine las dos. Nosotros seguimos este procedimiento.

```
# create new var
famsize <- sibsp + parch + 1
```

El siguiente paso consiste en analizar los valores que faltan para hacerse una idea de la exhaustividad de los datos:

```
> # NA analysis
> t.src.dim <- dim(t.src)
```

```
> nas <- apply(data.frame(t.src,famsize),2, function(i) sum(is.na(i)))
> nk <- 2
> data.frame(nas,ratio=round(nas/t.src.dim[1]*100,nk))
                             nas ratio
lfdall                         0   0.00
lfd                            0   0.00
fullname                       0   0.00
lastname                       0   0.00
title                          0   0.00
firstname                      1   0.05
secondfirstname              910  41.82
furthernames.relationships  1855  85.25
sex                            0   0.00
age                          263  12.09
sibsp                        867  39.84
parch                        867  39.84
ticket                         0   0.00
fare                           1   0.05
cabin                       1014  46.60
embarked                      10   0.46
boat                        1491  68.52
body                        1933  88.83
home.dest                    564  25.92
pclass                         0   0.00
survived                       0   0.00
Boarded                     1317  60.52
Position                       0   0.00
crew                           0   0.00
fate                           0   0.00
famsize                      867  39.84
```

La tabla 5.4 muestra el porcentaje de valores perdidos para las variables elegibles *sexo, edad, sibsp, parch, tarifa, camarote, embarcado, barco, cuerpo, pclass* y *superviviente*. Los elevados valores NA de *sibsp* y *parch* se deben a que no se dispone de datos de viajes familiares de los tripulantes. Es de suponer que sus familias no estaban a bordo, pero en sentido estricto esto no está claro con este conjunto de datos. A pesar de su supuesta magnitud, estos NA son bastante insignificantes y no se aplican a los pasajeros (ahí no hay NA), sino que se toman directamente después del tamaño de las familias. En cambio, los valores de NA muy elevados para *barco* y *cuerpo* deben tomarse en serio, ya que se trata de valores relevantes que faltan. Los valores del *bote* se deben a que los moribundos no solían estar presentes en los botes salvavidas y no era posible reconstruir a posteriori quién iba en cada bote. En cuanto al cuerpo, el problema es que sólo se recuperó una parte de los cadáveres y que muchos fueron arrastrados a las profundidades con el Titanic y desaparecieron irremediablemente. ¿Qué se puede reconstruir en tal caso? Una posibilidad es incluir los propios NA como parte de la tabla, es decir, como información (salida abreviada):

```
> # NA analysis of boat
> boat <- as.character(t.src$boat)
> boat[is.na(boat)] <- "NONE"
> boat <- as.factor(boat)
> str(boat)
Factor w/ 30 levels "?","1","10","11",..: 30 30 30 26 16 13 9 30 12 ...
> table(boat,survived)
survived
boat FALSE TRUE
? 0 6
1 0 12
10 0 33
...
D 1 22
NONE 1461 30
```

Tabla 5.4: *Variables y valores faltantes(NA – Titanic)*

Variable	NAs	% NAs
sex	0	0
age	263	12.09
sibsp	867	39.84
parch	867	39.84
fare	1	0.05
cabin	1014	46.60
embarked	10	0.46
boat	1491	68.52
body	1933	88.83
pclass	0	0
survived	0	0
famsize	867	39.84

Esto muestra que sólo 13 de las personas rescatadas en botes salvavidas registrados murieron, pero que hay 1461 NAs para survived=FALSE para los botes salvavidas. Por el contrario, sólo hay 30 NA para survived=TRUE para 672 personas que pueden asignarse claramente a los botes salvavidas. Interpretamos los NA para bote de tal manera que los NA probablemente sólo se refieren a las personas que murieron y nunca llegaron a subir a un bote. Esto se ve corroborado por la baja tasa de NA entre los rescatados de los que se desconoce la identidad. Para *body*, un análisis de este tipo no merece la pena, ya que sólo se refiere a los fallecidos, para los que sólo cabe la categoría "desaparecidos" en el caso de los NA. Por el momento, la interpretación es sencilla: con 243 cadáveres, sólo se recuperaron muy pocos de los 1474 fallecidos, el 16,5%. Llegado a este punto, resulta útil analizar si hubo una distribución equitativa de las clases (incluida la tripulación) en los botes salvavidas, una cuestión importante cuando se trata de la cuestión de la "equidad" en el rescate, si es que se puede utilizar este término en este contexto. Aquí lo utilizamos en el sentido de distribución equitativa de probabilidades. Para la tabla consideramos que la fila inferior contiene sólo los NA.

```
# analysis boat x pclass
table(boat, pclass)
```

Hay que añadir que las personas registradas en los botes salvavidas suman 672, pero el número total de personas rescatadas es de 702. Como no se puede suponer que las personas estuvieran continuamente en el agua helada durante horas hasta que fueron rescatadas en un barco, es necesario aclarar más la diferencia de 702 - 672 = 30 personas. En tal caso, lo mejor sería un análisis individual de cada caso para aclararlo. Se trata, pues, de personas que sobrevivieron pero que no pueden ser asignadas a un barco. Éstas pueden ser identificadas.

```
> # survived but no boat
> survived.noboat.ids <- which(survived == TRUE & boat == "NONE")
> length(survived.noboat.ids)
[1] 30
> survived.noboat <- t.src[survived.noboat.ids,]
> table(survived.noboat$pclass)
1st 2nd 3rd Crew
1   8   14  7
```

El procedimiento pretende inspirar cómo se pueden tratar los valores que faltan y cómo las comparaciones paso a paso dentro de los conjuntos de datos o entre ellos pueden eliminar parcialmente las ambigüedades o al menos proporcionar o al menos revelar información sobre cómo se produjeron los NA. Los datos mismos son, por supuesto, desconocidos.

Las NA pueden utilizarse por razones teóricas o estimarse a partir de un modelo, pero también pueden omitirse. Todo ello depende de la información contextual de que dispongamos y de las suposiciones que hagamos sobre el mecanismo de las NA. En el caso que nos ocupa, por ejemplo, podemos probar valores omitidos de *age* sobre una base argumentativa:

Conocemos características de las personas cuya edad no está clara, por ejemplo, el título y la clase (ambos sin NA) y posiblemente si la persona viajaba sola o con un grupo o familia. Esto nos permite distinguir entre niños/adolescentes y adultos y así acotar el rango de edad. Otro dato es que el 50% de las personas

```
quantile(age, na.rm =TRUE)
```

tienen entre 28 y 52 años, con una mediana de 40 años. Los cuantiles se pueden desglosar aún más, por ejemplo, por sexo y/o clase:

```
quantile(age[sex=="female"], na.rm=TRUE)
quantile(age[sex=="male"], na.rm=TRUE)
quantile(age[sex=="female" & pclass=="Crew"], na.rm=TRUE)
```

De este modo, se puede determinar una franja de edad para cada agrupación pertinente. Es más fácil hacer una distinción ad hoc con un diagrama de caja. Para facilitar la lectura del gráfico, las combinaciones *pclass* y *sex* se abrevian.

```
bp <- boxplot(age ~ pclass + sex, plot=FALSE)
str(bp)
bp$names
bp$names <- paste(rep(c("1st","2nd","3rd","C"),length(bp$names)/4),
            rep(c("f","m"),each=length(bp$names)/2),sep=":")
TITLE <- "Titanic"
SUB <- "age ~ pclass + sex"
bxp(bp, notch=TRUE, ylab="age", xlab="class vs. sex",
           main="", frame=FALSE,
           boxfill=rainbow(length(bp$names)/2), border=length(bp$names)/2)
mtext(TITLE, 3, line=2.5, cex=1.5)
mtext(SUB, 3, line=1, cex=1.1)
```

Los patrones de edad (véase la Fig. 5.21) de mujeres y hombres son muy similares dentro de sus grupos. Los de primera clase (por ejemplo, los hombres) son mayores que los de segunda o tercera. Las mujeres tienden a ser más jóvenes (excepto las tripulantes) como las mujeres de primera clase, pero con menos variabilidad. En el caso de los hombres, la edad de los tripulantes – más jóvenes que las mujeres – es aproximadamente la misma que la de los de segunda clase. No hay que olvidar los índices de base: en algunos casos, los números absolutos de casos difieren considerablemente (véase más adelante), por lo que una mirada a los porcentajes relativos puede, como siempre, inducir a error (véase el capítulo 6.3.4). Por ejemplo, n = 23 miembros de la tripulación son mujeres, mientras que m = 835 hombres pertenecen a la tripulación. La diferencia es de ¡≈36 veces!

```
> # base rates
> tab.class.sex <- table(pclass,sex)
> tab.class.sex[,2]/tab.class.sex[,1]
    1st     2nd     3rd      Crew
1.263889 1.669811 2.282407 36.304348
```

Según la proporción de valores perdidos subdivididos por sexo y clase y teniendo en cuenta ahora se podrían generar y utilizar valores de este espectro de los cuantiles anteriores del 25% al 75%.Otras variables se pueden tratar de forma equivalente. Esto ofrece la posibilidad de utilizar valores razonables para el subgrupo, pero desde luego no en relación con el individuo. Se trata simplemente de mantener estadísticas de grupo y eso rara vez es correcto para los individuos. Así que una desventaja es que las estadísticas existentes se mantienen relativamente constantes, por lo que si los datos disponibles distorsionan la población, la sustitución de las NA apoya esa misma distorsión.

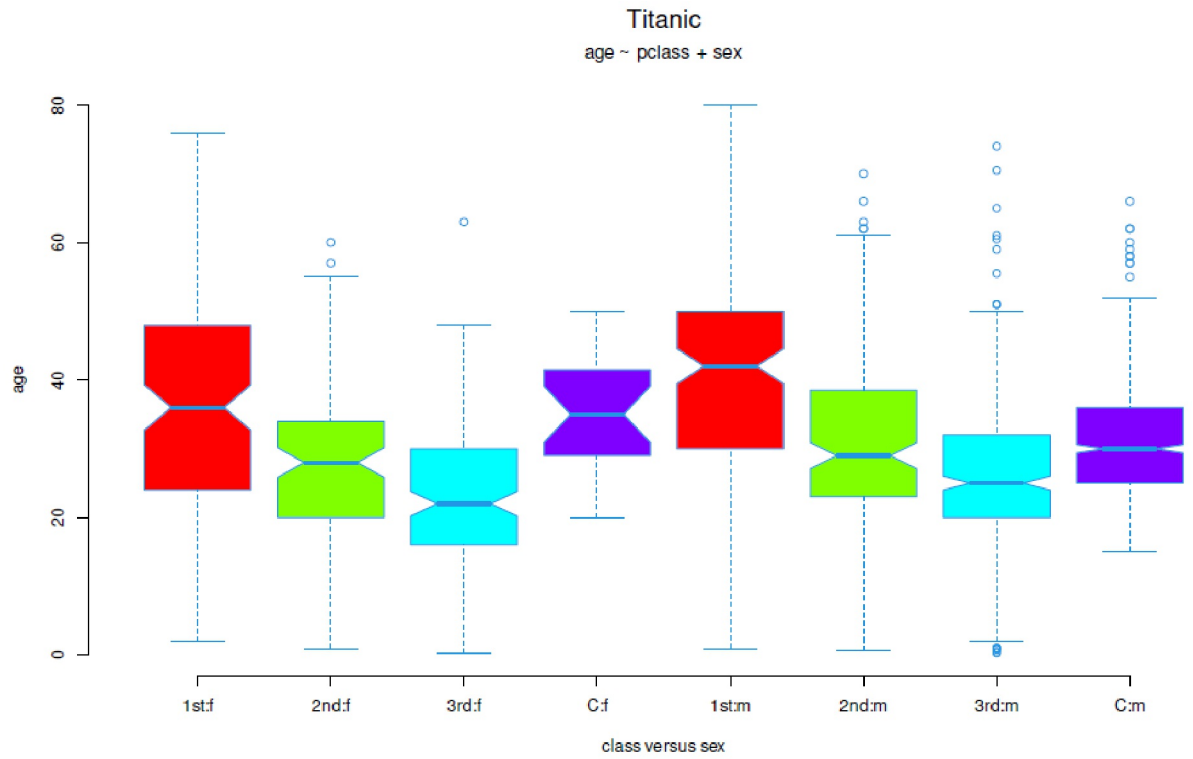


Figura 5.21. *Datos del Titanic (edad por sexo y clase, gráficos de caja)*

Sin embargo, si sólo se dispone de unos pocos valores y éstos apenas están relacionados con otras características personales, una sustitución es apenas posible, sobre todo si no se puede seguir ninguna característica numérica comprensible. Un ejemplo sería la variable *body* ya mencionada anteriormente, es decir la identificación de cadáveres recuperados en el Atlántico Norte.

```
> # NAs body
> sum(is.na(t.src$body))/length(t.src$body)
[1] 0.8883272
```

Aquí la tasa de NA es del 88.83%. Si bien podemos suponer que todas las personas fallecidas acabaron en el mar o murieron en un bote salvavidas – por tanto, el 100% – otra cosa es que flotaran en la superficie y fueran rescatados o no por los barcos de salvamento. Muchos fueron arrastrados por el Titanic y muchos probablemente simplemente fueron a la deriva con la corriente o fueron pasados por alto. Sustituir los NA aquí no tiene sentido y el beneficio tampoco es existente, ya que esta variable es prácticamente irrelevante para responder a nuestras preguntas iniciales.

Sin embargo, la sustitución de valores perdidos favorece una posición determinada y esperemos que bien fundamentada. Puede tratarse de una tendencia central (por ejemplo, la mediana) o de un modelo estadístico para mantener determinadas estadísticas de grupo (por ejemplo, con `mice()` del paquete `mice` de R). Siempre surge una incertidumbre adicional de los datos artificiales generados. Esta incertidumbre debe sopesarse frente al beneficio de poder calcular ahora con todos los datos. Nunca desaparece. Todas las conclusiones se basan en ella. Incluso la estadística bayesiana puede estimar modelos consistentes frente a las NAs (Jaynes, 2003), pero la ambigüedad fundamental sobre los valores perdidos permanece. No forman parte del análisis de los datos empíricos.

5.5.4.6.3 Creación de nuevas variables. Las preguntas anteriores equivalen a tabular variables y subgrupos entre sí para examinar las respectivas tasas de supervivencia desde distintos puntos de vista. Con fines orientativos, a veces tiene sentido crear nuevas variables a partir de otras ya existentes para hacer los datos más manejables y resumir características con pocos valores. Aquí algunas variantes:

5.5.4.6.3.1 Reducción del título. Si nos fijamos en la variable *fulltitle* (título completo), encontramos títulos que sólo aparecen en raras ocasiones y que son en parte específicos de cada sexo (salida abreviada):

```
> # reduce titles
> head(fulltitle)
[1] "Commander" "Lieutenant" "Lieutenant"
[4] "Sub-Lieutenant" "Mr" "Sub-Lieutenant"
> t(table(sex,fulltitle))
sex
fulltitle      female male
Capt           0      1
Col            0      4
...
Sub-Lieutenant 0      3
the Countess   1      0
```

Tiene sentido combinar algunos de estos títulos. Independientemente de la clase n = 6 títulos que parecen tener sentido. Se trata de *Lady, Master, Miss, Mr, Mrs* y *Sir*.

```
> upperclass.men <- c("Capt","Col","Commander","Don","Dr","Jonkheer",
+ "Lieutenant","Major","Rev","Sir",
+ "Sub-Lieutenant")
> upperclass.women <- c("Dona","Dr","Lady","the Countess")
> women.titles.young <- c("Miss","Mlle","Ms")
> women.titles.mature <- c("Mme","Mrs")
> men.titles <- c("Master","Mr") # we leave both as they are
> fulltitle.red <- fulltitle
> # before men due to "Dr" present in both strings
> fulltitle.red[fulltitle.red %in% upperclass.women] <- "Lady"
> fulltitle.red[fulltitle.red %in% upperclass.men] <- "Sir"
> fulltitle.red[fulltitle.red %in% women.titles.young] <- "Miss"
> fulltitle.red[fulltitle.red %in% women.titles.mature] <- "Mrs"
> table(fulltitle.red,sex)
sex
fulltitle.red female male
Lady           4      9
Master         0      61
Miss           275    0
Mr             0      1593
Mrs            210    0
Sir            0      24
> table(fulltitle.red,survived)
survived
fulltitle.red FALSE   TRUE
Lady          6       7
Master        30      31
Miss          86      189
Mr            1291    302
Mrs           44      166
Sir           17      7
```

5.5.4.6.3.2 Maternidad. Pueden optar las personas mayores de 18 años, que no viajen solas y sean mujeres. Hay que tener en cuenta que en ese momento, dependiendo de la clase clase, una persona podía ser considerada adulta a los 14 años, por ejemplo si tenía que trabajar. Pasamos por alto estos detalles, para ello tendríamos que hacer un análisis muy detallado del contexto histórico y de las personas implicadas y su pertenencia de clase. Por regla general, y no del todo correctamente, la edad de 18 años corresponde a la entrada en la edad adulta. Esta variable es útil para examinar más adelante si ser madre era una ventaja para la supervivencia.

```
> # create index possible mother
> table(sex == "female" & parch > 0 & age > 18)
FALSE TRUE
2039   97
```

Esto puede reducirse aún más eliminando a todas las personas con el título de *Miss*:

```
> table(sex == "female" & parch > 0 &
+ age > 18 & fulltitle.red != "Miss")
FALSE TRUE
2079   78
```

Esto da lugar a una nueva variable:

```
> possible.mother <- sex == "female" &
+ parch > 0 & age > 18 & fulltitle.red != "Miss"
> table(possible.mother,sex)
sex
possible.mother female male
FALSE              392 1687
TRUE               78  0
```

Como prueba se puede comprobar si la variable *possible.mother* es coherente, lo que debería ser a la vista de las selecciones lógicas.

```
> # cross-checks that variable was created properly
> sum(possible.mother[parch < 1],na.rm=TRUE)
[1] 0
> sum(possible.mother[age < 19],na.rm=TRUE)
[1] 0
> sum(possible.mother[fulltitle.red == "Miss"],na.rm=TRUE)
[1] 0
> sum(possible.mother[sex == "male"],na.rm=TRUE)
[1] 0
```

Todas las sumas dan cero, como debe ser.

5.5.4.6.3.3 Niños y ancianos. Este grupo resulta de una edad $<$ 18 años o una edad $>$ 60 años. En principio, los niños pueden subdividirse en bebés y niños muy pequeños, niños en edad escolar y adolescentes/pubescentes. Lo mismo se aplica a los mayores: lo dejaremos con una primera subdivisión simple. El procedimiento se mantiene constante.

```
> # create child index truth table
> child <- age < 18
> table(child,sex)
sex
child female male
FALSE 339 1400
TRUE  72  100
> # create senior index
> oldage <- age > 60
> table(oldage,sex)
sex
oldage female male
FALSE   404   1471
TRUE    7     29
```

5.5.4.6.3.4 Grupos de edad. En el caso de la edad, sería conveniente una ampliación de los índices existentes para niños y personas mayores. Son precisamente los grupos de edad que no tienen el mismo tamaño los que parecen interesantes para un análisis más profundo. Estos grupos de edad se sugieren a sí mismos: Niños de 0-18 años, adultos jóvenes de 19-30 años, adultos de 31-60 años y mayores $>$ 60 años. El comando de R `.bincode()` implementa esto. Los límites no se derivan de los datos en sí mismos, por lo que son concebibles otras agrupaciones, como adultos jóvenes y adultos o subdividir aún más a los niños.

En primer lugar, examinamos la distribución por edades:

```
> # age ranges
> # create asymmetric but content relevant age dimensions
> quantile(age,probs=seq(0,1,.1),na.rm=T)
0%      10%     20%     30%      40%     50%
0.1667 18.0000 21.0000 24.0000 27.0000 29.0000
60%      70%     80%     90%     100%
32.0000 35.0000 39.0000 45.0000 80.0000
> quantile(age,probs=seq(0,1,.05),na.rm=T)
0%       5%      10%     15%     20%      25%     30%
0.1667 13.0000 18.0000 20.0000 21.0000 22.0000 24.0000
35%      40%     45%     50%      55%     60%      65%
25.0000 27.0000 28.0000 29.0000 30.0000 32.0000 33.0000
70%      75%     80%     85%      90%     95%      100%
35.0000 37.0000 39.0000 42.0000 45.0000 52.0000 80.0000
```

Y ahora creamos los grupos de edad:

```
> # child - young adult - midterm + senior
> maxx <- max(age)
> age.bin.1 <- factor(.bincode(age, breaks=c(0,17,31,61, maxx),
           + right=TRUE, include.lowest=TRUE))
> levels(age.bin.1) <- c("child","youngadult","adult","senior")
> age.bin.1 <- factor(age.bin.1)
> table(age.bin.1,sex)
sex
age.bin.1 female male
child        72   100
youngadult  182  773
adult        150  604
senior       7    23
```

5.5.4.6.3.5 Tamaño de la familia y número de personas que viajan solas o en grupo. Esto supone un poco más de esfuerzo. El primer paso consiste en crear un índice familiar compuesto por el tamaño de la familia (véase más arriba *famsize <- sibsp + parch + 1*) combinando parejas/hermanos y padres/hijos. Se agrupan los hermanos y los padres/hijos. Esta variable puede ayudar a responder a la pregunta si era una ventaja no viajar solo y hasta qué tamaño de grupo existía esta ventaja.

```
> # create index family name + family size
> family.IDsize <- paste(lastname,famsize,sep=":")
> head(sort(table(family.IDsize),dec=TRUE))
family.IDsize
Sage:11 Andersson:7 Goodwin:8
     11           9         8
Smith:NA Asplund:7 Taylor:NA
      8          7         7
> tail(sort(table(family.IDsize),dec=TRUE))
family.IDsize
Yousif:1 Yousseff:1 Yrois:1
       1          1       1
Zanetti:NA Zarracchi:NA Zimmerman:1
        1            1            1
```

Varios sitios de análisis sugieren asignar personas a la variable *famsize* cuando las personas comparten cabañas pero no están registradas como miembros de la familia – como los siguientes casos.

```
> # create adjusted family size variable
> # for people sharing cabins but not registered as family members
> noccur <- data.frame(table(cabin))
> head(noccur)
    cabin Freq
1   A10      1
2   A11      1
3   A14      1
```

```
4   A16      1
5   A18      1
6   A19      1
> tail(noccur)
    cabin Freq
182 F33      4
183 F38      1
184 F4       4
185 G6       5
186 nopassenger 867
187 T            1
> dim(noccur)
[1] 187 2
```

De ellos, sólo interesan aquellos en los que se conocen tanto la letra de la cabina como el número. Por tanto, la longitud de la `string` debe ser mayor que uno.

```
> # remove if only letter is known, but no cabin number...
> noccur.cc <- subset(noccur, nchar(as.character(cabin)) > 1)
> dim(noccur.cc)
[1] 184 2
```

Y la cabina debe estar ocupada por más de una persona:

```
sharedcabins <- noccur.cc$cabin[noccur.cc$Freq > 1]
```

Ahora viene la comprobación para ver quién tiene una familia de un solo miembro, pero está registrado en un camarote compartido. A estos casos se les asigna un tamaño de familia de dos.

```
> # replace shared cabins but freq=1 by freq=2 for famsize.adj
> famsize.adj <- famsize
> sharedcabins.id <- which((famsize.adj == 1) &
+ (cabin %in% sharedcabins))
> sum(length(sharedcabins.id))
[1] 43
> # check whether really n=1
> sum(famsize.adj[sharedcabins.id] != 1)
[1] 0
> famsize.adj[sharedcabins.id] <- 2
> table(famsize,sex)
sex
famsize female male
    1   194    596
    2   123    112
    3   79     80
    4   29     14
    5   14     8
    6   10     15
    7   9      7
    8   3      5
    11  5      6
> table(famsize.adj,sex)
sex
famsize.adj female male
    1       171    576
    2       146    132
    3       79     80
    4       29     14
    5       14     8
    6       10     15
    7       9      7
    8       3      5
    11      5      6
```

Para la tripulación y para todos los NA fijamos ahora un tamaño de familia de uno, porque menos no es posible y uno es el valor mínimo correcto. Luego sigue la tabla correspondiente y la comprobación si todavía hay Nas.

```
> # for staff use family size = 1 (=one, i.e. alone with oneself)
> length(famsize.adj[pclass=="Crew"])
[1] 858
> is.na.ids <- which(is.na(famsize.adj))
> famsize.adj[is.na.ids] <- 1
> table(famsize.adj)
famsize.adj
 1    2   3   4  5  6  7 8 11
1614 278 159 43 22 25 16 8 11
> sum(is.na(famsize.adj))
[1] 0
```

Ahora creamos el índice de tamaño de la familia. Las categorías son *viajeros solos, parejas, grupos de de 3 a 4 personas* y *grupos más grandes* o *familias con más de 4 personas*.

```
# create small family size index
maxx <- max(as.numeric(names(table(famsize.adj))))
# 0 | 1 alone | 2 duo | group 3+4 | >=5
travelno.bin.1 <- factor(.bincode(famsize.adj, breaks=c(0,1,2,4,maxx),
           right=TRUE, include.lowest=TRUE))
levels(travelno.bin.1) <- c("alone","duo","group3+4","group>4")
travelno.bin.1 <- factor(travelno.bin.1)
table(travelno.bin.1,sex)
```

5.5.4.6.3.6 Tarifas. Estas (Fowler, 2019, véase también el debate en Quora, 2017) oscilaban para el Titanic de 3 a 8 libras (tercera clase) a 12 libras (segunda clase) y 30 libras (primera clase). La Suite Parlour costaba en aquella época la enorme cantidad de 875 libras. Si se convierte eso al valor monetario de hoy en día son 172,- $ a 460,- $ (tercera clase), 690,- $ (segunda clase), 1724,- $ (primera clase) y 50 000 $ (suite de salón). De aquí se pueden derivar categorías que corresponden aproximadamente a la variable *pclass*. A la tripulación se le asigna su propia categoría. Lo primero que hay que hacer es examinar los propios datos antes de transformarlos.

```
# drop crew who did not pay but was paied
fare.p <- fare[fare > 0]
describes(fare.p)
quantile(fare.p,probs=seq(0,1,.1),na.rm=T)
```

Ahora viene la clasificación a la nueva variable *fare.bin.1*.

```
# bins: 0 - 0.7854 - 10.5 - 21.679 - 39.688 - 512.329
maxx <- max(fare, na.rm=TRUE)
fare.bin.1 <- factor(.bincode(fare, breaks=c(0,1,8,12,30,maxx),
right=TRUE, include.lowest=TRUE))
levels(fare.bin.1) <- c("Crew","Third","Second","First","Suite")
fare.bin.1 <- factor(fare.bin.1)
table(fare.bin.1)
table(fare.bin.1,sex)
table(fare.bin.1,pclass)
```

Como muestra la última tabla *fare.bin.1* contra *pclass*, hay diferencias:

```
        pclass
fare.bin.1  1st 2nd 3rd  Crew
Crew         10  12   4   858
Third        1   0    342 0
Second       0   44   139 0
```

```
First       66  186  170 0
Suite       249 41   53  0
```

Esto se debe a que los límites tendrían que ser investigados con mucha precisión en cuanto a dónde comienza cada clase y termina. Así, una tabulación de tarifas muestra bastantes tarifas diferentes.

```
> # number of categories
> length(table(fare))
[1] 281
```

Este es uno de los casos en los que es necesario el trabajo manual y la verificación de cada dato individual. Las razones de las discrepancias no están claras. Puede que los padres pagaran por el conjunto de sus familias y su tarifa fuera tan elevada en total que pudiera clasificarse en otra clase, pero que en realidad se produjera como consecuencia de la tarifa pagada en grupo. Además, existen posibles destinos diferentes a los que se podría haber llegado con diferentes tarifas. ¿Cuántas personas no pagaron nada y no formaban parte de la tripulación? De las personas no categorizadas como tripulación, hay que restar otras 9 personas que pertenecían al grupo de asegurados, eran en cierto modo parte de la tripulación, pero viajaban en camarotes de primera y segunda clase.

```
> sum(fare[pclass != "Crew" & t.src$crew != "guarantee.group"] == 0,
            + na.rm=TRUE)
[1] 17
> sum(fare[child == TRUE] == 0, na.rm=TRUE)
[1] 18
> sum(fare[child == TRUE] != 0, na.rm=TRUE)
[1] 154
```

Así, para 7 niños, la tarifa es cero, pero para 74 niños, hay un tarifa mayor de cero. Por tanto, nuestra tesis es correcta sólo en parte. Además, hay 17 - 7 = 10 adultos que también tienen un cero y no pertenecen a la tripulación. Por lo tanto, la tabla *fare.bin.1* debe utilizarse con precaución, ya que no está claro qué contiene realmente.

Es concebible la derivación de otras variables. Por ejemplo, podría crearse un índice a partir de los apellidos y (ajustado, véase más arriba) el tamaño de la familia y dotarlas de un índice para reducir el número de categorías y distinguirlas de los viajeros no familiares. A partir de ahí, se obtienen índices de supervivencia específicos para las familias.

Ahora vienen los análisis exploratorios propiamente dichos, que no se publican en su totalidad. Corresponde al lector examinarlos más detenidamente. En aras de la simplicidad, hay análisis tabulares y gráficos, que en la práctica van juntos.

5.5.4.7 Análisis tabulares

En el capítulo 4.4.14.1 se examinó la paradoja de Simpson y se hizo referencia a la importancia de los porcentajes base a la hora de interpretar las fuerzas relativas y los porcentajes. Así, cuando se dice, la "probabilidad de supervivencia de hombres (o mujeres o niños) era mayor o menor que xyz %", debe ir seguida inmediatamente de la pregunta: "¿Cuáles fueron los porcentajes de base?" Un caso límite fictico demuestra la necesidad. Por ejemplo, entre la tripulación del Titanic podría haber un solo niño. Pero ignorar los índices de base conduce entonces a las dos afirmaciones siguientes, lógicamente VERDADERAS, que proporcionan poca información para la realidad:

- Si el niño no se salva, todos los miembros de la tripulación que eran niños no pudieron salvarse y murieron (= 100 %).
- Si el niño se salva, todos los miembros de la tripulación que eran niños pudieron salvarse completamente y sobrevivieron (= 100 %).

Con una tasa base de n = 1, las afirmaciones porcentuales sin tasas base absolutas carecen de sentido. Pero en principio, no son afirmaciones falsas. Sólo la información adicional sobre la tasa base revela el

significado. Si sólo hubo un niño y sobrevivió, la segunda frase es correcta. Si, por el contrario ese niño murió, la primera frase es correcta. ¿Se puede deducir algo de todo esto? No mucho. Con mucho más de n = 2200 pasajeros, a una persona se le asigna el factor 1/2200, es decir, 0.045%.

Si todo el mundo en el Titanic hubiera sobrevivido excepto este único niño, la tasa de mortalidad seguiría siendo del 0:05% y, sin embargo, el único niño a bordo habría muerto - no menos trágico. Si el Titanic fuera un pequeño velero con n = 10 personas, una sola persona -por ejemplo, el niño en cuestión- ya representa el 10% de todas las personas a bordo. Las proporciones numéricas adquieren su significado del contexto, que por tanto debe elegirse sabiamente para no difundir información distorsionadora. Sin esta información adicional, específica de la situación, no debe interpretarse ningún dato sobre la calidad de la misma. Esto es así no sólo cuando se interpretan los índices de supervivencia, sino en general. Si, por ejemplo, se dice que "se salvaron el doble de hombres (en términos absolutos) que de mujeres", eso no significa que las mujeres tuvieran menos posibilidades de sobrevivir (véase el estudio de caso de las tasas de admisión en la Universidad de Berkeley, es decir, la paradoja de Simpson, véase el capítulo 4.4.14.1). Esto sólo es cierto si como máximo viajaban a bordo el doble de hombres que de mujeres. La afirmación cambia inmediatamente si la información adicional dice: "En total, había tres veces más mujeres que hombres a bordo" (entonces las mujeres tendrían una probabilidad demasiado pequeña) o "Sólo había la mitad de mujeres que de hombres a bordo" (entonces las mujeres tendrían una probabilidad demasiado pequeña) o "Sólo había la mitad de mujeres que de hombres a bordo" (en ese caso, las probabilidades serían justas y se distribuirían por igual). Con la información adicional, las proporciones cambian, al igual que las números absolutos de mujeres, hombres y niños supervivientes en el contexto general.

Es importante recordar esto para las próximas tablas, de lo contrario se extraerán conclusiones erróneas. extraídas. A continuación esbozamos las posibilidades y no las cubrimos todas combinatoria y completamente. Comenzamos con un análisis sociodemográfico detallado de las personas a bordo – los tipos de base antes mencionados. Por razones de espacio y para evitar el aburrimiento no los imprimimos todos. Claro, con el código R se pueden imprimir todas.

```
> # table analyses
> structable(age.bin.1 ~ pclass + sex,
+ split_vertical = c(TRUE, TRUE, FALSE, FALSE))
age.bin.1      child youngadult adult senior
pclass  sex
1st     female  8       44        75      6
        male    7       38        98      12
2nd     female  18      52        33      0
        male    16      81        62      5
3rd     female  46      78        27      1
        male    60      191       95      3
Crew    female  0       8         15      0
        male    14      63        349     3
```

Esto se ve muy bien formateado, pero es difícil calcular con él (¡desgraciadamente!) cuando intentamos aplicar, por ejemplo tratamos de aplicar `addmargins()` o `apply()` a la tabla, así que `table()` y `addmargins()`, lo que no parece mucho peor.

```
addmargins(table(pclass,sex))
addmargins(table(age.bin.1,sex))
addmargins(table(embarked,sex))
```

Repetimos lo mismo para las tasas de supervivencia de los subgrupos respectivos. Merece la pena disponer de una tabla maestra que contenga todas las variables relevantes.

```
# survival
mastertable <- table(pclass, age.bin.1, sex, survived)
ftable(mastertable)
```

`margin.table()` es sólo una envoltura de `apply()`, por lo que lo siguiente da resultados idénticos:

```
# is identical to
mastertable
apply(mastertable,c(1,2,3,4),sum)
```

Lo mismo vale para la suma de todas las frecuencias en la tabla:

```
# identical to
margin.table(mastertable)
sum(mastertable)
sum(apply(mastertable,1,sum))
```

De esto se puede deducir las sumas de sobrevivientes en los sub-grupos:

```
> # summed over var 1 in the order of appearance str(mastertable)
> margin.table(mastertable,1) # pclass
pclass
1st 2nd 3rd Crew
288 267 501 855
> margin.table(mastertable,2) # age.bin.1
age.bin.1
child youngadult adult senior
172     955       754   30
> margin.table(mastertable,3) # sex
sex
female male
411    1500
> margin.table(mastertable,4) # survived
survived
FALSE TRUE
1282  629
```

Lo mismo es apropiado para combinaciones de variables

```
# summed over var 1 and then var 2
ftable(apply(mastertable,c(1,4),sum)) # pclass x survival
ftable(apply(mastertable,c(2,4),sum)) # age.bin.1 x survival
ftable(apply(mastertable,c(3,4),sum)) # sex x survival
```

y para sumas en tablas multidimensionales:

```
# summed over var 1 and then var 2 and then var 3
ftable(apply(mastertable,c(1,2,4),sum)) # pclass x age.bin.1 x survival
ftable(apply(mastertable,c(1,3,4),sum)) # pclass x sex x survival
```

Para las tasas de supervivencia recurrimos a prop.table(), de nuevo una envoltura, esta vez para `sweep()`. En este sentido, las siguientes llamadas son idénticas. Primero limitamos los decimales

```
> # survial rates, no more sums
> op.orig <- options(digits=2)
> tx <- table(sex,survived)
> MARGIN <- 1
> sweep(tx, MARGIN, margin.table(tx, MARGIN), "/", check.margin=FALSE)
survived
sex     FALSE TRUE
female  0.27  0.73
male    0.80  0.20
> sweep(tx, MARGIN, apply(tx, MARGIN,sum), "/", check.margin=FALSE)
survived
sex     FALSE TRUE
female  0.27  0.73
male    0.80  0.20
> prop.table(table(sex,survived),m=1)
survived
```

```
sex     FALSE TRUE
female  0.27  0.73
male    0.80  0.20
```

Para simplificar, utilizamos `prop.table()`. Una llamada sin especificar la opción por defecto margin (abreviada con m = NULL) mostrará valores relativos en relación con toda la tabla. Pasar la opción margen sólo tiene efecto si se utiliza más de una variable. Evaluación por filas o columnas:

```
> prop.table(table(sex))
sex
female male
0.22   0.78
> prop.table(table(sex,survived))
survived
sex     FALSE TRUE
female  0.06  0.16
male    0.62  0.16
```

Las opciones m = 1 y m = 2 dan las proporciones para las filas y las columnas respectivamente. Una breve comprobación muestra que es correcto.

```
prop.table(table(sex,survived),m=1)
apply(prop.table(table(sex,survived),m=1),1,sum)
prop.table(table(sex,survived),m=2)
apply(prop.table(table(sex,survived),m=2),2,sum)
```

A partir de aquí, *todas las demás* tablas se pueden derivar *combinatoriamente*.

```
# further tables
# analysis based on two variables
prop.table(table(sex,survived))
prop.table(table(sex,survived),m=1)
prop.table(table(sex,survived),m=2)
prop.table(table(pclass,survived))
prop.table(table(pclass,survived),m=1)
prop.table(table(pclass,survived),m=2)
prop.table(table(age.bin.1,survived))
prop.table(table(age.bin.1,survived),m=1)
prop.table(table(age.bin.1,survived),m=2)

# analysis based on three variables
ftable(prop.table(table(pclass,sex,survived)))
# pclass x survival
ftable(prop.table(table(pclass,sex,survived),m=1))
# sex x survival
ftable(prop.table(table(pclass,sex,survived),m=2))

# analysis based on two variables - but table contains four variables
# pclass x survival
ftable(prop.table(table(pclass,age.bin.1,sex,survived),m=c(1,4)))
# age.bin.1 x survival
ftable(prop.table(table(pclass,age.bin.1,sex,survived),m=c(2,4)))
# sex x survival
ftable(prop.table(table(pclass,age.bin.1,sex,survived),m=c(3,4)))

# analysis based on three variables - but table contains four variables
# pclass x age.bin.1 x survival
ftable(prop.table(table(pclass,age.bin.1,sex,survived),m=c(1,2,4)))
# pclass x sex x survival
ftable(prop.table(table(pclass,age.bin.1,sex,survived),m=c(1,3,4)))
# age.bin.1 x sex x survival
ftable(prop.table(table(pclass,age.bin.1,sex,survived),m=c(2,3,4)))

# more difficult to understand, but contrasts each possible
```

```
# sub-combination for survival TRUE vs. FALSE
# pclass x age.bin.1 x sex
ftable(prop.table(table(pclass,age.bin.1,sex,survived),m=c(1,2,3)))
```

El punto de partida para responder a las hipótesis relativas a las ventajas para las madres y los viajeros individuales frente a los viajeros en grupo figura en las tablas siguientes:

```
> # possible mothers
> table(possible.mother)
possible.mother
FALSE TRUE
2079   78
> prop.table(table(possible.mother))
possible.mother
FALSE TRUE
0.964 0.036
> prop.table(table(possible.mother,survived))
survived
possible.mother FALSE  TRUE
FALSE           0.6727 0.2911
TRUE            0.0079 0.0283
> prop.table(table(possible.mother,survived),m=1)
survived
possible.mother FALSE TRUE
FALSE           0.70  0.30
TRUE            0.22  0.78
> prop.table(table(possible.mother,survived),m=2)
survived
possible.mother FALSE TRUE
FALSE           0.988 0.911
TRUE            0.012 0.089
```

Podemos repetirlo para el tamaño de grupo:

```
# traveling alone versus in group
table(famsize.adj)
prop.table(table(famsize.adj))
prop.table(table(famsize.adj,survived))
prop.table(table(famsize.adj,survived),m=1)
prop.table(table(famsize.adj,survived),m=2)
```

Por último, volvemos a la configuración por defecto para la visualización de los decimales.

```
# reverse options (digits)
options(op.orig)
```

La salida de las tablas podría automatizarse mediante un recuento de conjuntos de potencia (Gürtler & Oldenbürger, 2005). Un conjunto de potencias contiene todos los subconjuntos de combinaciones posibles. Para el caso anterior de m = 4 variables, resultarían las siguientes combinaciones:

```
# power set count R-Code
m <- 4
potlist <- list()

for(i in 1:m) potlist[[i]] <- combn(m,i)
potlist
```

En la salida es importante la distinción entre el primer y el último elemento de la lista. La primera entrada enumera las cuatro variables individualmente (como un vector, una variable por entrada), mientras que la última entrada enumera las cuatro variables juntas en paralelo (como una columna de una tabla).

```
str(potlist)
```

Volviendo a las tablas, rápidamente se hace evidente que son confusas y que requieren la creación de una tabla completamente nueva para cada entrada adicional. También es necesario familiarizarse con las tablas antes de poder utilizarlas de forma comprensible, la diferencia entre valores absolutos y relativos se hace evidente y se establece una comprensión intuitiva general. Los análisis gráficos de los mismos datos ofrecen una alternativa a las tablas.

5.5.4.8 Análisis gráficos

Los paquetes `vcd` y `vcdExtra` de R ofrecen una variedad de opciones de análisis gráfico para mostrar y explorar datos categóricos, especialmente en el contexto de modelos log-lineales. Utilizamos `doubldecker()` y `mosaic()`. Ahora debería quedar claro en las tablas anteriores que las variables de los datos del Titanic interactúan entre sí y sólo adquieren un significado real en el contexto de las demás. A un gráfico de mosaico se le da la anchura o la altura de los bloques respectivos en función de las frecuencias relativas relacionadas con la categoría asociada (dimensión) y sus subdivisiones. Además, es posible codificar por colores los residuos, es decir, la diferencia entre las frecuencias observadas frente a las esperadas, sobre la base de un modelo log-lineal. Todo esto es posible con `mosaic()` del paquete `vcd` de R. Utilizamos esta información, pero no la relativa a las significaciones, ya que aquí se practica el AED y no la estadística confirmatoria. Resaltar las desviaciones en las frecuencias celulares esperadas debido a las tasas de base parece ser una opción interesante para comprender mejor las estructuras subyacentes. Las desviaciones con respecto a las frecuencias esperadas no tienen nada que ver con las pruebas, sino que inicialmente sólo son una afirmación sobre las clasificaciones y motivo de otras consideraciones. No consideramos las pruebas de significación reales del modelo log-lineal. Incluso en el caso de una validación estadística clásica del modelo, la opción de marcar con colores es muy útil para interpretar un gráfico de mosaico. `doubledecker()`, en cambio, no permite derivaciones estadísticas inferenciales y sólo puede utilizarse de forma exploratoria. Este tipo de gráfico visualiza la dependencia de una variable categórica (típicamente binaria codificada como estado de supervivencia en este caso) de otras variables categóricas. Las figuras 5.22 y 5.23 contienen los gráficos respectivos y deben interpretarse de forma relativa, es decir, no absoluta. Para una mejor legibilidad de las expresiones de muchas variables – los valores respectivos de *superviviente, sexo* e *hijo* se abrevian en las nuevas variables *survived.TF, sex.FM* y *child.AC*.

```
# graphical analyses
survived.TF <- factor(survived, labels=c("F","T"))
sex.FM <- factor(sex, labels=c("F","M"))
child.AC <- factor(child, labels=c("A","C"))
doubledecker(survived.TF ~ pclass + sex.FM + child.AC,
           gp=gpar(fill=c("violetred3","greenyellow")) )
mosaic(~ survived.TF + pclass + sex.FM + child.AC, pop=FALSE,
           shade=TRUE, legend=TRUE,
split_vertical=c(TRUE,TRUE,FALSE,FALSE),gp=shading_hcl,
           gp_args=list(h=c(130,43), c=100, l=c(90,70)))
```

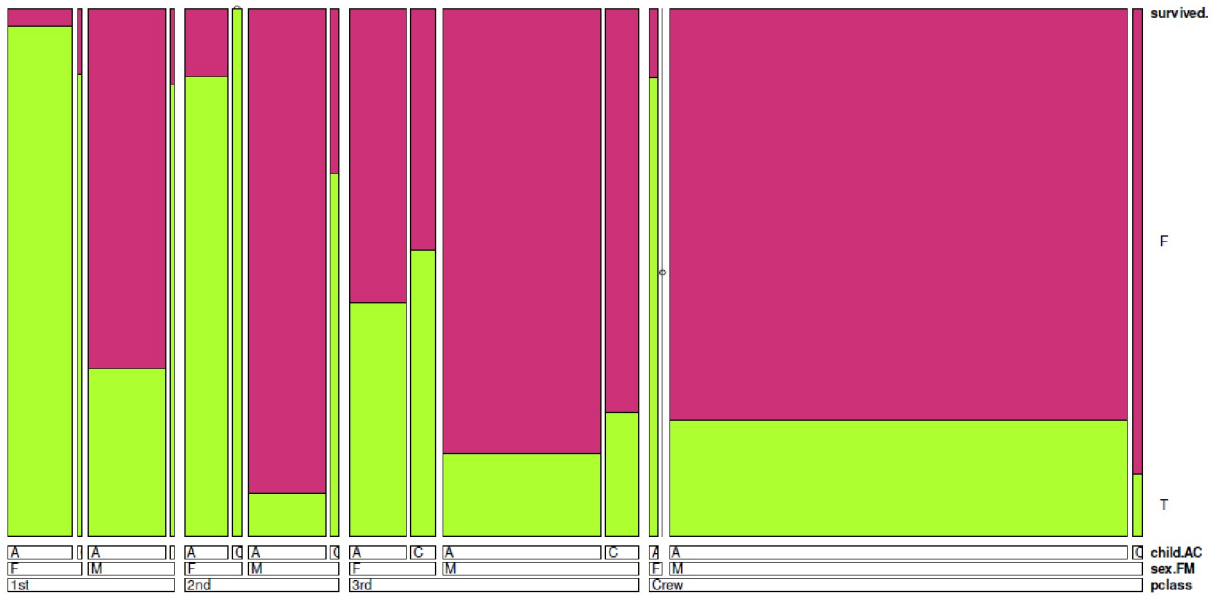


Figura 5.22: *Titanic (Tasas de sobrevivencia, gráfico doubledecker/biplano)*

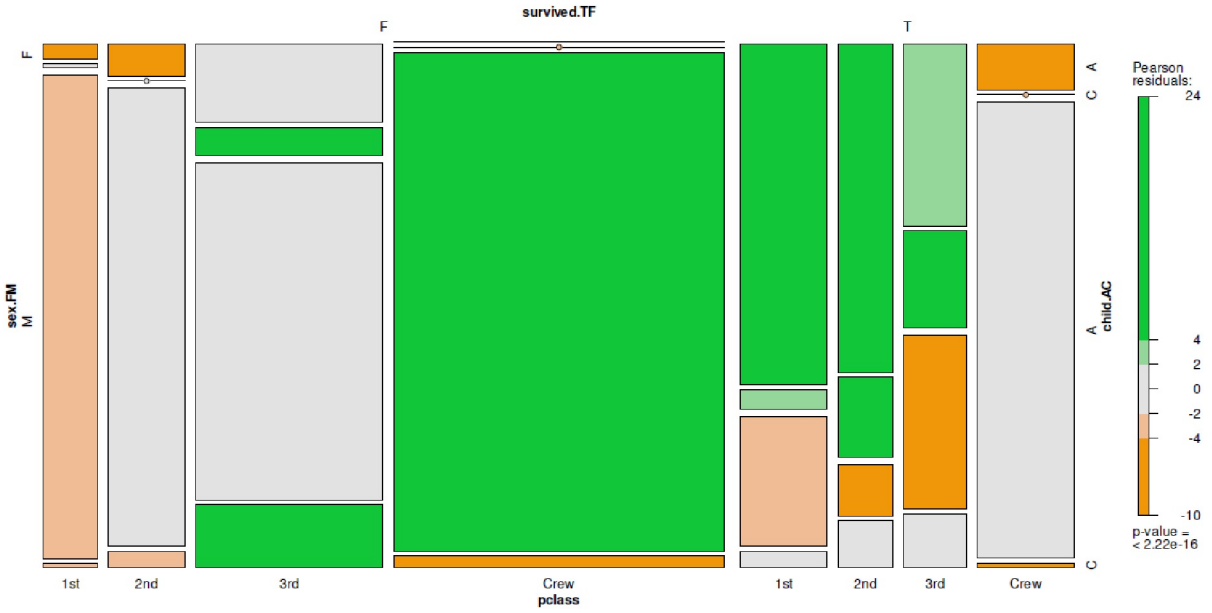


Figura 5.23: *Titanic (Tasas de sobrevivencia, gráfico mosaico)*

También merece la pena observar la distribución de la supervivencia en función de la edad (véase la Fig. 5.24) y la tarifa (véase la Fig. 5.25. La tarifa se traza sin los miembros de la tripulación. Ad-hoc hemos creado una función R alrededor de `hist()` llamada `hist.titanic()` para trazar los datos en función de la variable sobrevivido. Esta no es una función general, pero es una solución rápida y sucia para el uso diario. Utilizamos funciones de R que utilizamos muy a menudo y las adaptamos fácilmente a otras condiciones. Esto no es perfecto, pero es pragmático, porque ahorra muchas llamadas individuales de comandos R y da el gráfico directamente con una llamada.

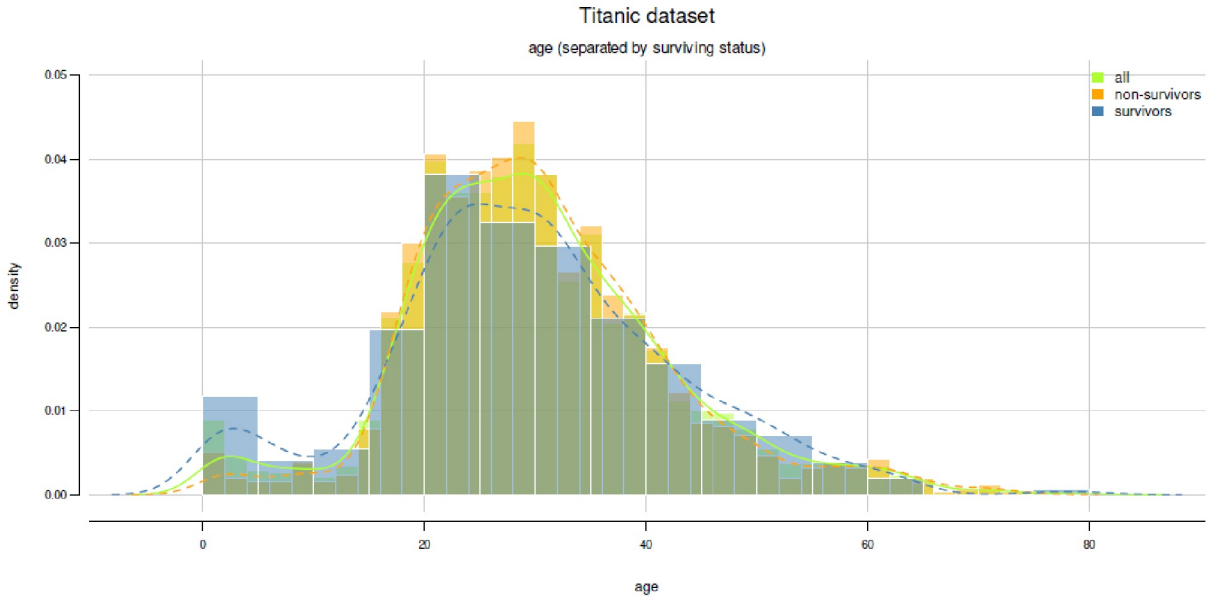


Figura 5.24: *Titanic (Edad, histograma)*

```
# hist plot
hist.titanic(daten=age, TITLE="Titanic dataset",
           SUB="age (separated by surviving status)",
           xaxtext="age")
hist.titanic(daten=fare.p, TITLE="Titanic dataset",
           SUB="fare (separated by surviving status)",
           xaxtext="fare.p (fare > 0)")
```

Como el gráfico de tarifas está muy distorsionado por las suites muy caras, repetimos el gráfico para las tarifas inferiores a 100 £ (véase la Fig. 5.26), lo que equivale a ampliar la imagen y simplemente ocultar la gama por encima de 100 £.

```
hist.titanic(daten=fare.p[fare.p <100], TITLE="Titanic dataset",
           SUB="fare (separated by surviving status)",
           xaxtext="fare.p (0 < fare < 100)")
```

Esto hace que los datos parezcan un poco más legibles. Globalmente, la variable edad no parece tener una influencia directa en la supervivencia. Esto no significa que la edad no esté mediada por género y la clase social.

Como complemento a los gráficos de mosaico anteriores, es posible escribir frecuencias u otros valores en las celdas respectivas. Esto es muy elegante, ya que combina la salida concreta de una tabla con el enfoque intuitivo de los gráficos. De este modo, se puede también mostrar los residuos.

```
#add values to cells/ tiles
mosaic(~ survived.TF + pclass + sex.FM + child.AC, pop=FALSE,
           labeling=labeling_residuals,
           shade=TRUE, legend=TRUE, gp=shading_hcl,
           gp_args=list(h=c(130,43), c=100, l=c(90,70)))
```

Además se podría ordenar según hipótesis interesantes para nosotros:

```
mosaic(~ survived.TF + pclass + sex.FM + child.AC, pop=FALSE,
           labeling=labeling_residuals,
           shade=TRUE, legend=TRUE,
           split_vertical=c(TRUE,TRUE,FALSE,FALSE),gp=shading_hcl,
           gp_args=list(h=c(130,43), c=100, l=c(90,70)))
```

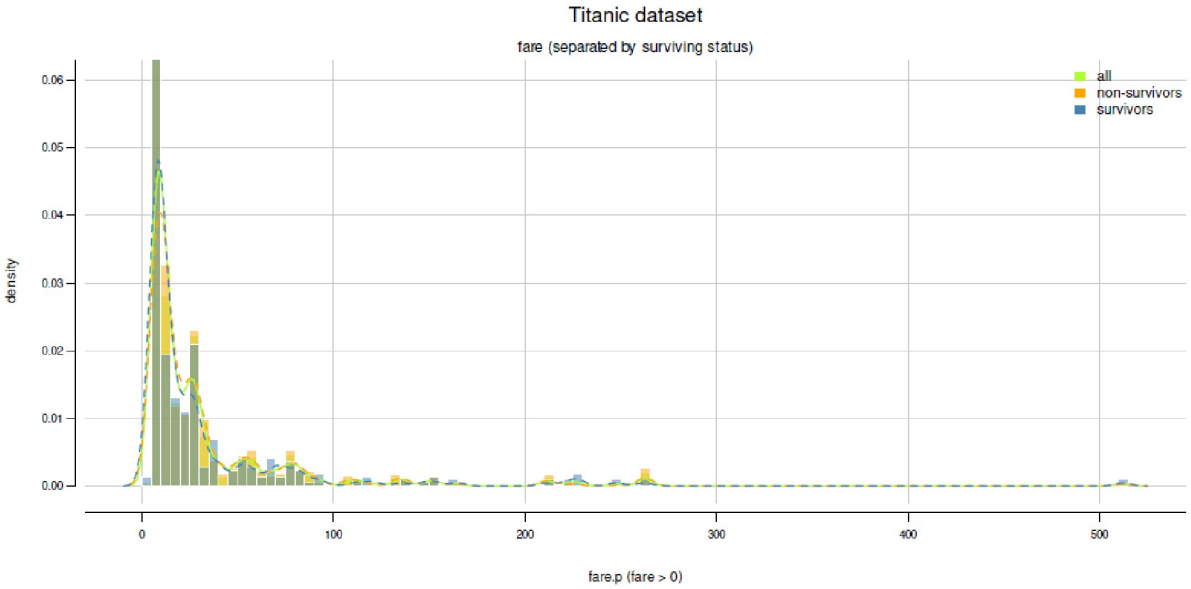


Figura 5.25: *Titanic (Tarifa, histograma)*

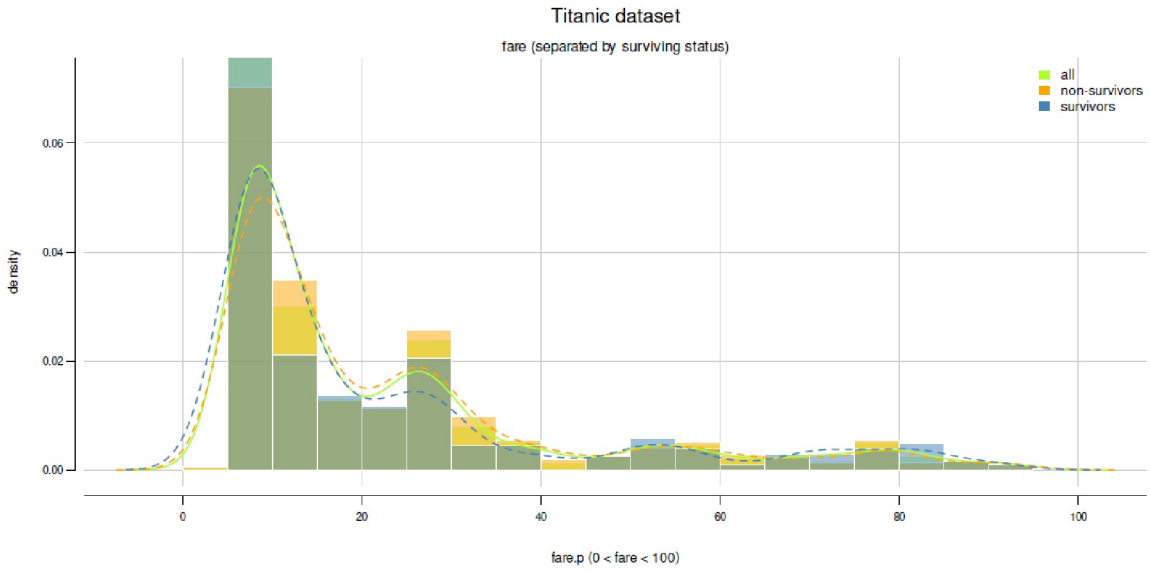


Figura 5.26: Titanic *(Tarifa $<$ £100, histograma)*

Como alternativa, las células pueden controlarse directamente con una función separada. Comenzamos con las frecuencias simples.

```
age.bin.1.CYAAS <- factor(age.bin.1, labels=c("C","YA","A","S"))
mtab <- table(pclass, age.bin.1, sex.FM, survived.TF)
mosaic(mtab,pop=FALSE)
labeling_cells(text=mtab, margin=0)(mtab)
```

Y esto funciona para variables y estadísticas seleccionadas, no sólo para frecuencias. Ahora combinamos el gráfico de mosaico con las tasas de supervivencia (véase la Fig. 5.27).

```
stab <- structable(survived.TF ~ child.AC + sex.FM + pclass)
stab
tab.prop.sex <- round(prop.table(table(pclass,sex.FM,survived.TF),
                  m=c(2,3))*100,nk) #sex
ftable(tab.prop.sex)
mosaic(stab, pop=FALSE, shade=TRUE, legend=TRUE, gp=shading_hcl,
                  gp_args=list(h=c(130,43), c=100, l=c(90,70)), split=TRUE)
labeling_cells(text=tab.prop.sex)(stab)
```

El código R anterior crea una visión tabular, contenida en el gráfico de mosaico, de cómo se distribuyen las tasas de supervivencia (lógicamente TRUE frente a falso) dentro de los géneros y entre las clases. El gráfico de mosaico respectivo debe adaptarse con la llamada `structable()` de forma que la salida gráfica y la organización parezcan significativas e intuitivamente comprensibles. La figura 5.27 debe interpretarse de forma que los valores de *sex.FM, survived.TF* e independientemente de *child.AC* puedan interpretarse como una distinción entre las respectivas expresiones de *pclass*. La tabla `tab.prop.sex` contiene la información pertinente. Así, los valores para *survived.TF* = Y y *sex.FM* = F son para las clases 1ª = 38.72%, 2ª = 26.18%, 3ª = 29.53%y tripulación = 5.57% – es decir, mujeres que sobrevivieron, separadas por clase social. Sin embargo, los porcentajes no deben interpretarse sin las correspondientes tasas básicas. En el gráfico, estos valores se muestran tanto en la esquina superior izquierda como en la esquina inferior izquierda, ya que la variable *child.AC* no se incluyó numéricamente como diferenciación adicional. Tal diferenciación adicional es posible – o la adición de o sustituirla por otras variables como *age.bin.1*, embarcado o posible.madre. Entonces habría que crear de nuevo la tabla `tab.prop.sex`, seguida de la llamada a `mosaico()`. Encontrará ejemplos de combinaciones interesantes de tablas en la sección anterior sobre análisis tabulares.

Tarea 5.2: Titanic: Creación de otros gráficos

La tarea para los lectores consistiría en crear otros gráficos interesantes utilizando las variables anteriores. Compara tus resultados con los análisis que se pueden encontrar en Internet o la viñeta del paquete `vcd` de R. Puede, por ejemplo, crear histogramas y estimaciones de densidad para subgrupos diferentes y combinados - por ejemplo, para sexo, sexo combinado con con el estado de supervivencia o adicionalmente para las divisiones de clase.

¡Tenga en cuenta las tasas base!

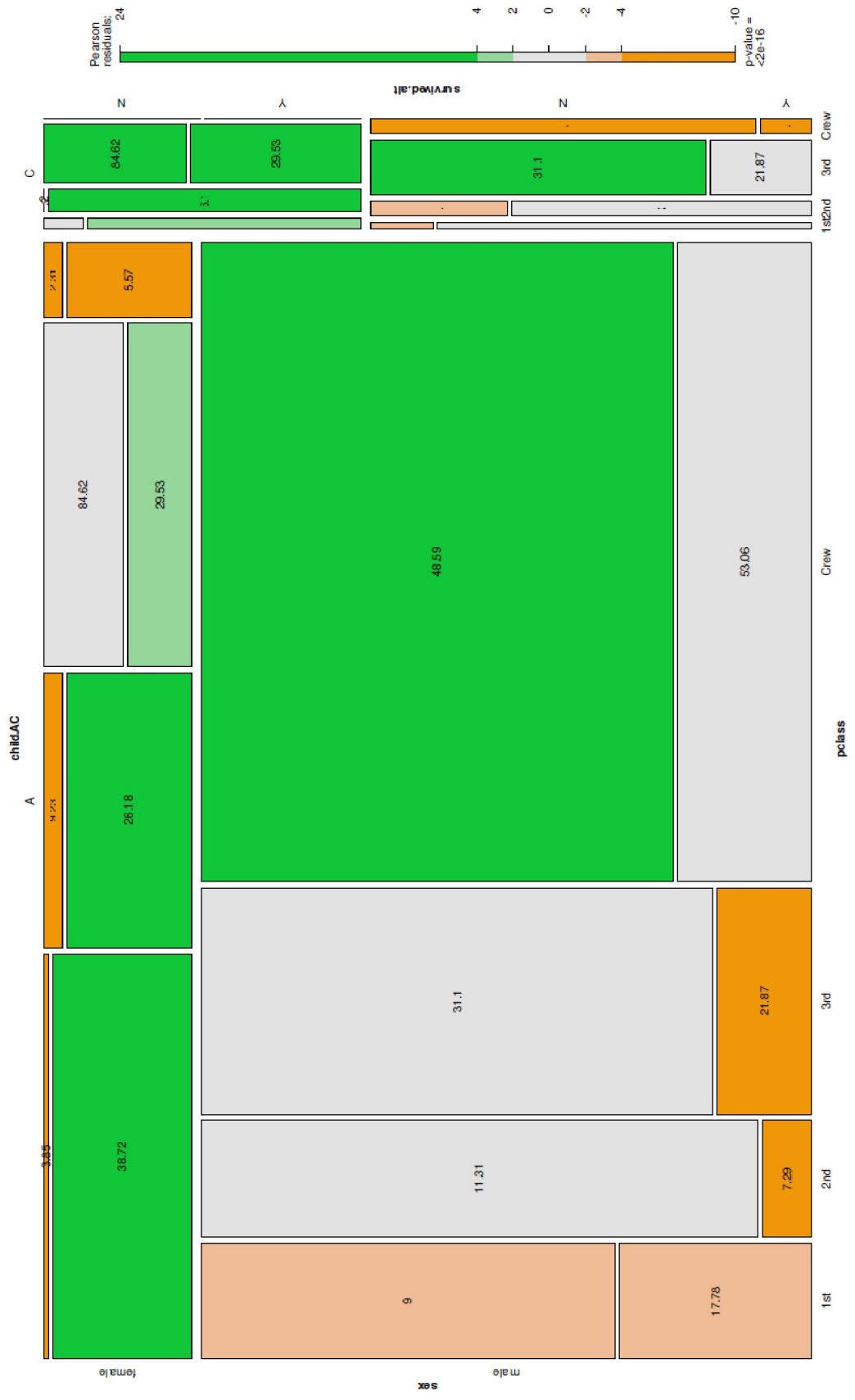


Figura 5.27: *Titanic (Tasas de sobrevivencia con números, gráfico mosaico)*

5.5.4.9 Discusión de los análisis exploratorios sobre Titanic

Las siguientes consideraciones se basan en los anteriores análisis exploratorios tabulares y gráficos de los datos del Titanic. Éstos no son completos y asumimos que nuestra conclusión inicial contiene, por tanto, errores. Los lectores interesados pueden tomar nuestros análisis como punto de partida para mejorarlos y ampliarlos.

El *análisis tabular* muestra que murieron demasiadas personas, y muchas más de las que cabría esperar dada la falta de botes salvavidas. Sólo se salvó el 32.2%. Esto significa que murieron 2/3 de todas las personas a bordo del Titanic.

```
> # tables for final discussion
> prop.table(table(survived))
survived
FALSE      TRUE
0.6773897  0.3226103
```

Lo que sigue se aplica a las variables individuales:

Grupo – Si bien el grupo de mujeres, con un 22.5% de las personas a bordo, forma un porcentaje de mujeres entre los rescatados del 51.1%. Dentro del grupo de mujeres, el 73.4% fueron rescatadas, mientras que sólo el 22.3% de los hombres. Ser mujer fue probablemente una ventaja.

Clase – En cuanto a la clase, se observa que el 28.5% de las personas rescatadas pertenecían a la primera clase y el 28.8% a la tripulación. Sin embargo, la tripulación representaba el 39.4% del total de personas;
la primera clase sólo representaba el 15%. Viajar en primera clase también era una ventaja.

Edad – En cuanto a las estructuras de edad, dentro de cada grupo de edad, los niños representaban el mayor porcentaje de supervivientes, con un 44.6%. Es decir que de todos los niños, se salvó el 44.6%. En el caso de los jóvenes y los adultos sólo el 31.6% y el 30.3% respectivamente. Los ancianos se salvaron el 41% de las veces. Dentro del grupo de los rescatados, los niños representaron el 8%. Como era de esperar, los adultos constituyeron la mayor proporción de personas rescatadas, con un 52.6%. Los adultos jóvenes representaban el 23% y los ancianos el 16.4%. Por otra parte, el grupo de los niños representaba sólo el 5.9% de todas las personas del Titanic. Para los adultos jóvenes es el 24%, para los adultos 57% y para los ancianos 13.1%. Ser niño era una pequeña ventaja a bordo del Titanic, o ser una persona mayor. Lo que hay que examinar más detenidamente es si esta afirmación se relativiza al añadir la clase social.

Maternidad – ser madre también ayudaba a sobrevivir. Aunque aparentemente las madres de todas las personas, el 78.5% de las madres se salvaron y el 9% de los supervivientes eran madres.

Viajar solo o en grupo – Viajar solo o en grupo parece un poco más complejo. Un grupo de 4 personas mostró el mayor factor de impacto, que denotamos como la frecuencia relativa de sobrevivir dividida por la frecuencia relativa de estar en el Titanic (= tasa base relativa). Esto no tiene en cuenta el hecho de que la tripulación puede también puede dividirse en grupos y que estos son grupos ad hoc de amistad suelta y, debido al trabajo común, de interdependencia.

```
> # travelers alone vs. group
> fsize.tab <- prop.table(table(famsize.adj))
> fsizeXsurv.tab <- prop.table(table(famsize.adj,survived),m=2)
> fsize.tab
famsize.adj
1           2           3           4           5
0.741727941 0.127757353 0.073069853 0.019761029 0.010110294
6           7           8           11
0.011488971 0.007352941 0.003676471 0.005055147
```

```
> fsizeXsurv.tab
survived

famsize.adj FALSE      TRUE
1           0.816824966 0.584045584
2           0.082089552 0.223646724
3           0.046811398 0.128205128
4           0.008819539 0.042735043
5           0.010854817 0.008547009
6           0.013568521 0.007122507
7           0.008141113 0.005698006
8           0.005427408 0.000000000
11          0.007462687 0.000000000
```

Un factor de impacto de exactamente uno significa que proporcionalmente tantos fueron rescatados como proporcionalmente estaban representados en el barco. Un valor superior a uno significa que se salvaron más de los esperados y un valor inferior a uno significa lo contrario: murieron más de los esperados. Así pues, los grupos de 4, 3 y 2 personas tenían ventaja, ya que todos sobrevivieron más de lo esperado. Los grupos de 5 personas o los que viajaban solos tenían menos probabilidades de sobrevivir de lo esperado, seguidos de los grupos de 7 y 6 personas. Los grupos grandes de 8 y 11 personas estaban prácticamente ausentes en el barco, pero seguían teniendo las probabilidades relativas más bajas. Habría que complementar el análisis con datos como la proporción de rescatados dentro de cada clase de viaje en comparación con la absoluta.

Con esta primera impresión, lo dejamos en este punto. Por supuesto, el mundo no es tan simple como las variables *sex, pclass, age.bin.1*, *possible.mother* y *famsize.adj* sugieren por sí solas. Siendo realistas, cabe esperar muchas interacciones entre las variables. Éstas pueden analizarse mediante otras tablas, análisis gráficos o estadística inferencial mediante complejos modelos jerárquico-lineales. Sin embargo, el procedimiento permanece constante en principio – sólo se hacen más complejas las tablas, que son más fáciles de examinar mediante un análisis gráfico. A continuación pasaremos a ello. El análisis gráfico con `doubledecker()` (véase la Fig. 5.22) lo interpretamos como sigue. Para simplificar las cosas, la probabilidad de supervivencia se abrevia como *SR* (survival rate; tasa de supervivencia). La variable *SR* es una medida relativa.

- En el caso de las mujeres, indica un descenso de la SR de la primera a la segunda y a la tercera clase. Entre la tripulación, las mujeres apenas estaban representadas. No obstante, su SR debe considerarse muy alta con un número absoluto bajo.
- En general, los hombres tienen una SR más baja que las mujeres. En la primera clase, sin embargo, su SR era superior que en las demás clases o en la tripulación.
- En cuanto a los niños, cabe señalar que su RE fue mayor en las niñas que en los niños y bastante alta en la primera clase, menor en la segunda y significativamente peor en el tercero. Las clases más bajas tenían más niños, sin embargo, que estaban mal representados en general. Los niños de la tercera clase tenían aproximadamente el mismo nivel bajo de SR que los adultos de la segunda clase, lo que suponía una desventaja.
- Globalmente, las mujeres tenían mayor SR que los hombres; y los niños tenían mayor SR que los adultos, aunque su índice de base era mucho más bajo.
- Las mayores pérdidas absolutas se produjeron entre los hombres, concretamente en la tripulación y en la tercera clase. Sin embargo, en términos absolutos, éstos también formaban los grupos con las mayores tasas básicas absolutas.

El *análisis gráfico del mosaico* mediante `mosaic()` completa el cuadro – consideramos explícitamente la marca de color de los residuos (residuos de Pearson) como la diferencia de lo observado frente a la esperada, escalada a la desviación estándar estimada. En la Figura 5.28, las desviaciones hacia arriba (> 0) se marcan con verde y las desviaciones hacia abajo (< 0) con naranja.

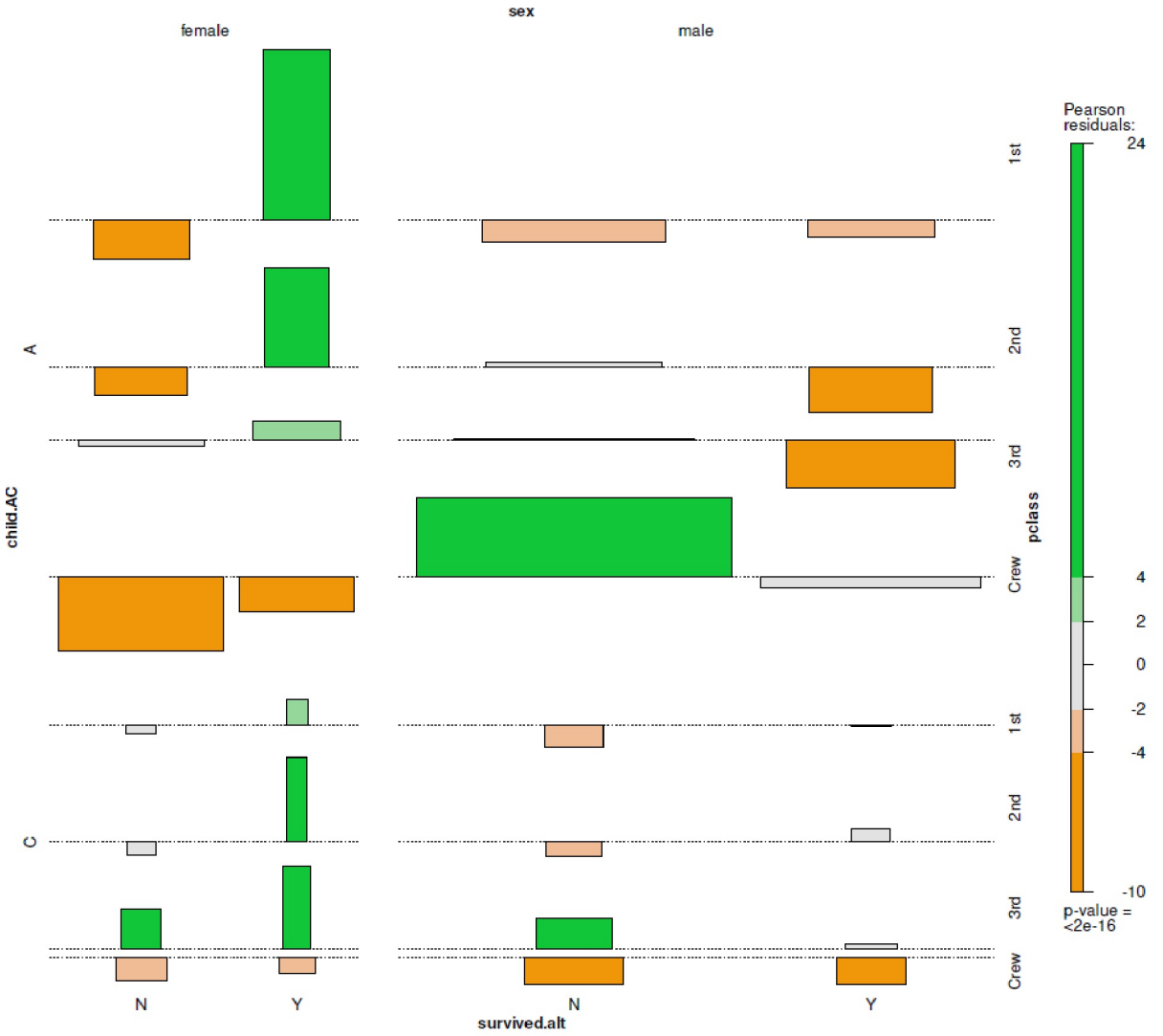


Figura 5.28. *Titanic (tasas de supervivencia con números, assocplot)*

- En el caso de survived = FALSE hay dos distinciones.
  1. Residuales inferiores a cero (= murieron menos o sobrevivieron más de lo esperado). Esto se aplica a las mujeres (1ª y 2ª clase, así como a la tripulación) y, en menor medida, a los hombres (1ª clase) y niños como miembros de la tripulación (niñas y niños).
  2. Residuales superiores a cero (= murieron más o sobrevivieron menos de lo esperado). Esto incluye a los hombres de la tripulación y, en menor medida, a los niños (niñas y niños de 3ª clase).

- En el caso de survived = TRUE hay dos distinciones.
  1. Residuales menores a cero (= han muerto más o han sobrevivido menos de lo esperado). Los casos positivos son hombres (2ª y 3ª clase), mujeres de la tripulación y – de modo atenuado – hombres (1ª clase) o niños como miembros de la tripulación (niñas y niños).
  2. Residuales superiores a cero (= sobrevivieron más o murieron menos de lo esperado). Se trata de mujeres (1ª y 2ª clase), niñas (2ª y 3ª clase) y – de modo atenuado – mujeres (3ª clase) o niñas (1ª clase).

En general, observamos que la clase baja – equivalente a un estatus socioeconómico más bajo – suponía un mayor riesgo de muerte, independientemente del sexo y la edad. Entre clases, el género y la edad moderaban. En general, las mujeres sobrevivieron mejor, a menos que permanecieron en la tercera clase.

Dentro de la categoría de hombres, los de la primera clase tenían una mayor ventaja en la primera clase con respecto a las demás. Los niños tenían una clara ventaja en las dos primeras clases, pero ésta ya no era efectiva en la tercera.

La mayor ventaja de supervivencia parecía ser para los niños y las mujeres en las clases superiores, y para los varones en la primera clase. Si volvemos a nuestras pegadizas tesis anteriores (véase el capítulo 5.5.4.2), podemos responder a nuestras preguntas de la siguiente manera sobre la base de la información exploratoria actual, es decir, nuestro estado de error actual:

1. El lema "las mujeres y los niños primero" se aplicó efectivamente en el Titanic, pero con un peso considerablemente menor cuanto más baja era la clase. En la clase más baja esta ventaja ya no existía realmente. Esto se refiere a la probabilidad general de supervivencia y no dice nada sobre la rapidez y cómo subieron realmente las mujeres y los niños a bordo de los botes salvavidas.
2. Si hubieras podido elegir quién eras, las mujeres y los niños (excepto los de tercera clase) o los hombres de primera clase eran los favoritos para sobrevivir, quienes no hubieras querido ser eran los hombres y especialmente los de tercera clase o los miembros de la tripulación (números absolutos). En términos relativos, sin embargo, dentro del grupo de hombres, sobrevivieron más tripulantes que hombres de tercera clase o de segunda clase (los menos).
3. El Titanic no era un buen patrón. Una enorme proporción de la tripulación no fue rescatada y murió. El Titanic tampoco era una buena compañía de cruceros: demasiada gente murió innecesariamente. Las razones residen en el diseño técnico y no sólo en los procedimientos específicos de la noche del desastre. Los problemas son sistémicos. Por ejemplo, el número de botes salvavidas era demasiado reducido. Además, hay una alta probabilidad de conducción irresponsable - conducir a 22,5/23 nudos corresponde a > 97 % de la velocidad máxima y esto con al menos seis avisos de iceberg esa noche, que obviamente no fueron escuchados. Desde luego, no se ignoraba que los icebergs también están presentes bajo la superficie del mar y que, por tanto, pueden embestir a un barco desde abajo, por lo que no basta con no embestirlos por encima del nivel del mar. Hoy en día, es una norma internacional que un buque pueda detenerse a toda velocidad en un radio de 15 esloras. Un Estado puede ampliar esta longitud hasta un máximo de 20 esloras de barco en casos excepcionales. Si aplicamos esto al Titanic, con una eslora de poco menos de 270 m, obtendremos unos 5,4 km si el Titanic hubiera iniciado una parada de emergencia, cosa que no hizo. Las maniobras de parada de emergencia suponen tal esfuerzo para el sistema de propulsión que incluso en los superpetroleros actuales sólo se practican en casos de extrema urgencia. Si se extrapola, una parada de emergencia de un petrolero gigante de hasta 450 m tarda un máximo de 30 minutos. Esto deja más claro lo pronto que deberían haber actuado el capitán y los navegantes del Titanic. Ya que, como se ha mencionado, el Titanic navegaba por delante a máxima potencia en esas condiciones. En nuestro caso, sin embargo, no se trató de un problema de parada, sino de un error de navegación. El radio de giro de los grandes buques actuales es de un máximo de cinco esloras, para hacernos una idea de la flexibilidad de los grandes buques.
4. Los patrones muestran que, efectivamente, hubo efectos globales – por ejemplo, el estatus socioeconómico, la edad o el sexo, pero estos factores se influyen mutuamente. Un análisis meticuloso de los procesos objetivamente reconstruibles y una mayor información contextual, que vaya más allá de las listas de pasajeros y tripulantes disponibles, permitirían hacer afirmaciones más claras sobre los factores de influencia interactivos. El cálculo de modelos log-lineales jerárquicos para la estimación precisa de los efectos principales y efectos de interacción o árboles de regresión para la clasificación (paquete R `rpart`) con conjuntos de datos de entrenamiento y validación sería útil de aquí en adelante para comprobar un modelo estadístico más preciso a lo largo del trabajo preliminar. Este procedimiento ulterior pasaría de exploratorio a inferencial y los supuestos aquí formulados podrían comprobarse con mayor precisión. La agrupación de variables podría servir de base para modelos multinivel.

Una vez más, es significativo que las cifras absolutas de casos y las probabilidades relativas de supervivencia en el contexto de distintas variables no parezcan decir lo mismo al principio y se contradigan

entre sí. Las diferencias sólo se ponen de manifiesto mediante el análisis, algo que se sabe desde el estudio de la Universidad de Berkeley sobre la presunta discriminación de las mujeres para acceder a la universidad (Bickel, Hammel & O'Connell, 1975, véase el capítulo 4.4.14.1 sobre la paradoja de Simpson y este estudio de caso). Nos parece aún más importante considerar por igual tanto el número absoluto de casos como las probabilidades relativas. Ambos responden a preguntas diferentes y no se contradicen en absoluto. Utilizando el ejemplo de la catástrofe del Titanic, las cifras absolutas de los casos nos dan una idea de cómo se relacionan los factores individuales con la supervivencia, si sólo se examinan paralelamente ellos y no otros factores. Una tasa de base elevada (= número de personas de una determinada categoría a bordo) corresponde a una expectativa correspondiente en cuanto a la probabilidad de supervivencia y a la expectativa de "condiciones justas", es decir, que la probabilidad de supervivencia depende únicamente de la proporción relativa de la frecuencia con la que un grupo de personas estaba representado en el barco. Si un grupo de personas estaba representado con mucha frecuencia (por ejemplo, los tripulantes varones), cabe esperar un número mayor tanto entre los supervivientes como entre los fallecidos que entre un grupo que sólo estaba ligeramente representado (por ejemplo, los niños). Si las expectativas se desvían de la realidad, por ejemplo porque un grupo de personas sobrevivió con especial frecuencia o no, comienza el análisis, que incorpora entonces factores contextuales para determinar las razones de las desviaciones sobre una base relativa. Por ejemplo, las probabilidades relativas muestran que los miembros de la tripulación tenían más posibilidades de rescate que los de tercera y segunda clase, pero inferiores a los de primera clase. Es decir, dentro del grupo de los que sobrevivieron, había sobre todo muchas personas de primera clase y de la tripulación, aunque en términos absolutos la mayoría de los muertos eran de tercera clase y de la tripulación, simplemente porque viajaba mucha menos gente en primera clase. Así pues, cabe señalar que, en términos absolutos, el número de personas que viajaban en primera clase era significativamente mayor que el de las que lo hacían en segunda, por lo que la proporción relativamente pequeña de la segunda clase dentro del grupo de supervivientes en comparación con la primera clase es sorprendente. Por otra parte, si nos fijamos en la probabilidad relativa de supervivencia dentro de cada clase en lugar de entre clases, la probabilidad de supervivencia fue menor en la segunda clase que en la primera clase, pero significativamente mayor que en la tercera clase y la tripulación. Desde esta perspectiva, una clase superior era ventajosa. La tabla 5.5 resume estos diferentes puntos de vista.

Tabla 5.5: *Titanic (diferentes perspectivas de supervivencia, `pclass`)*

Clase	Número de casos		Tasa de sobreviviencia	
	absoluto	relativo	entre las clases	dentro de la propia clase
1a	326	0.15	0.29	0.61
2a	283	0.13	0.17	0.42
3a	709	0.33	0.26	0.26
Tripulación	858	0.39	0.29	0.24
Σ	2176	1	1	–

Tarea 5.3: Preguntas y afirmaciones sobre el Titanic

La tarea para los lectores consistiría en comparar estos resultados con las tesis planteadas al principio del estudio de caso. ¿Qué ha resultado cierto, qué no y sobre qué no podemos hacer ninguna afirmación razonable a pesar del análisis?

Con esto concluye por el momento el análisis de los datos del Titanic. Debería haber quedado claro que un proceso AED es complejo para llegar a conclusiones razonables, pero no permite conclusiones causales causalizar o comprobar rigurosamente las conclusiones estadísticas. Dependiendo de los datos y de la información contextual disponible esto puede irse de las manos, como muestra el ejemplo del Titanic. Pero a cambio se obtienen afirmaciones útiles y comprobables para posteriores análisis. Una prueba confirmatoria es difícil, ya que obviamente no se desea la replicación y desastres de barcos de este orden afortunadamente no están disponibles. Por lo tanto, los análisis trabajan con una división de la muestra en dos mitades: una muestra de entrenamiento y una muestra de prueba. Con los datos del el primer grupo se construye el modelo y se prueba con los datos del segundo.

En el capítulo 12, dedicado al análisis de implicantes, se examina de nuevo el Titanic en el contexto de un análisis comparativo cualitativo según Ragin (1987), con el fin de identificar los casos y las configuraciones de código asociadas, es decir, escenarios típicos de conjuntos de variables, que conducen a la supervivencia o la muerte en el Titanic desde un punto de vista lógico. No obstante, nos parece impresionante las hipótesis y posibles conclusiones que pueden extraerse del análisis paralelo de información contextual, información descriptiva en forma tabular o gráfica, y el sentido común.

Una vez concluidas nuestras consideraciones sobre la catástrofe del Titanic, continuaremos con los estudios empíricos publicados con especial atención a las decisiones tomadas a lo largo del AED, algunos de los cuales se basan en métodos mixtos.

5.5.5 Liderazgo en contextos educativos

En un estudio español realizado por Gento, Huber, González, Raúl, Palomares & Orden (2015b) sobre el tema del comportamiento de liderazgo en contextos educativos, se utilizaron ante todo cuestionarios (escala de 9 puntos) para evaluar la relevancia e importancia del tema del liderazgo y los complejos temáticos asociados en el contexto correspondiente. Gento (2001a, 2001b, 2002) describe las características críticas del comportamiento de liderazgo en contextos educativos (carismático, emocional, expectante/ anticipatorio, profesional, contributivo/ participativo, cultural, formativo, administrativo). La base de datos fue n = 1027 cuestionarios cumplimentados de España, Letonia y un número menor de ocho países latinoamericanos. La importancia de estas ocho dimensiones fue de 7.31 a 7.66 en la escala de 9 puntos, mientras que las puntuaciones de presencia o evidencia fueron ligeramente inferiores, de 6.82 a 7.33 (Gento, 2014). Estas puntuaciones globales bastante altas sugerían la importancia general de esta cuestión. Sin embargo, es difícil detectar directamente patrones subyacentes en los datos para comprender más concretamente la cuestión. Esto sería necesario para poder utilizar los conocimientos así adquiridos en la práctica, es decir, en la formación y el perfeccionamiento profesional. Al mismo tiempo, los patrones subyacentes son políticamente relevantes, por ejemplo, para un debate sobre los cambios estructurales necesarios en las estructuras de dirección de las instituciones educativas de forma concreta y basado en una argumentación concreta y empíricamente fundamentada.

Sobre la base de un diseño de profundización en el contexto de los métodos mixtos (Mayring, 2001), se realizaron entrevistas cualitativas semi-estructuradas sobre los temas del *comportamiento del liderazgo en la educación* y la *influencia del liderazgo en la sostenibilidad y la calidad de las instituciones educativas*. Los entrevistados fueron educadores en liderazgo que ya habían respondido al cuestionario. El objetivo de estas entrevistas era de conocer los puntos fuertes y débiles del comportamiento del liderazgo, la influencia sobre la educación en general y las instituciones en particular, así como las influencias contextuales y los modos. Se trata de un diseño cuantitativo-cualitativo secuencial que se basa en el diseño para cartografiar de forma más exhaustiva toda la situación informativa.

Se realizaron N = 32 entrevistas, que se codificaron e interpretaron según el paradigma de codificación (véase el capítulo 9). Por un lado, los resultados confirmaron los de los cuestionarios. Por otro lado, llamaron la atención sobre detalles interesantes para comprender mejor los valores de importancia, de presencia y evidencia. A pesar de todo, seguía siendo difícil obtener una verdadera imagen global y descubrir las estructuras subyacentes dentro de las entrevistas y entre las entrevistas. Llegados a este punto, se decidió adoptar un diseño de Conversión 3.0 (Huber, Gürtler y Gento, 2018) mediante el análisis de los datos del

contexto interpretativo (codificación) utilizando métodos del AED. En concreto, se exploraron cuantitativamente las propiedades de los códigos. Para una mayor claridad, los códigos de las ocho dimensiones enumeradas se condensaron en categorías más abstractas de mayor alcance (meta-códigos en AQUAD7, Huber & Gürtler, 2012, sección 7.3). Esta agregación produjo códigos relativamente claros y significativos que contenían un nivel de abstracción suficiente para una comparación entre individuos. Dado que se trata de entidades discretas se decidió utilizar un método discreto, a saber, el análisis jerárquico de conglomerados (HCA) y un Escalado Multidimensional (MDS) para visualizar la proximidad y la distancia de las categorías entre sí en el espacio bidimensional y tridimensional. Con frecuencias enteras que, como mínimo, son de escala de intervalo y, en principio, pueden ser de naturaleza continua. Los códigos o entrevistas constituyen las unidades reales de análisis y se tratan como unidades discretas. Las dimensiones resultantes figuran en la tabla 5.6 con sus abreviaturas en inglés.

Tabla 5.6: *Estudio de Gento et al. (2015b, categorías y abreviaturas)*

Abbrev.	Categoría	Abbr. en inglés
D1	carismático	charisma
D2	emocional	emotion
D3	expectante/previsor	anticipation
D4	profesionalidad	professionalism
D5	participante	participation
D6	cultural/contexto	culture / context
D7	formativo (training, perfeccionamiento)	edu/training
D8	administrativo	administration

Para la formación de las distancias, se calcularon en primer lugar las *distancias euclidianas* simples, es decir, el camino directo. Desde nuestro punto de vista, otras medidas de distancia (por ejemplo, Manhattan, Mahalanobis; Handl, 2002) son por supuesto posibles, pero deben estar bien justificadas. No se nos ocurre ninguna buena razón para no utilizar el modelo simple de la geometría euclidiana, que corresponde al teorema de Pitágoras en el espacio bidimensional. La distancia euclidiana es *una métrica* y satisface la *desigualdad del triángulo* (Oldenbürger, 1994). Es un caso especial de la métrica de Minkowski y, por tanto, está estrechamente relacionada con la métrica de Manhattan o métrica de "city blocks" (Handl, 2002), que no sigue la vía directa directo, sino que va "a la vuelta de la esquina" (es decir, en ángulo recto) para medir una distancia – como el distrito neoyorquino de Manhattan, que se dividió en cuadrados siguiendo el modelo de la ciudad de Mannheim dividida en cuadrados, en el sur de Alemania. El cuadrado de la distancia euclidiana es un caso especial del cuadrado de la distancia Mahalanobis (Mahalanobis, 1936). Para la distancia euclídea de los puntos p y q con las coordenadas px, py, qx y qy se aplica, pues, en el espacio bidimensional

$$d(p_{xy}, q_{xy}) = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2} \tag{5.1}$$

y para la métrica Manhattan en el caso especial de dos dimensiones se aplica la ecuación

$$d(p_{xy}, q_{xy}) = |p_x - q| + |p_y - q_y| \tag{5.2}$$

Para el HCA, en general nos limitamos a unas pocas combinaciones de procedimientos de formación de distancias y aglomeración e intentamos no aumentar innecesariamente la complejidad ya existente. El carácter exploratorio, pero casi arbitrario, de la combinación de métodos de formación de distancias y aglomeración se hace especialmente evidente con los métodos de agrupación jerárquica. Hay muchos procedimientos tanto para la formación de distancias como para la aglomeración (Bock, 1974). Si éstos se combinan entre sí, se crea un inmenso mar de combinaciones de distancia-aglomeración, de modo que la preferencia por una combinación en un caso concreto apenas puede justificarse de forma significativa en un caso individual concreto, ni tampoco puede justificarse de forma inequívoca desde el punto de vista estadístico. Y mucho menos esta abundancia de combinaciones permite comprobar sin ambigüedades si se ha utilizado el modelo correcto. Por ello, hemos decidido limitarnos a unos pocos procedimientos habituales (véase el cuadro 5.7) y a los que entendemos en el caso. En consecuencia, preferimos comparar entre sí dos o tres soluciones de aglomeración diferentes a nivel de contenido, por ejemplo, en qué medida surgen diferencias entre las soluciones y dónde están los puntos en común. A menudo preferimos esta comparación al supuesto uso de la "única combinación de procesos correcta". No podemos comparar todas las combinaciones y, en nuestra opinión, eso sería extralimitarse de todos modos. Posiblemente complicaría innecesariamente las cosas sencillas y orientaría el debate hacia los análisis en lugar de hacia los contenidos investigados. Esto oscurecería lo esencial. En su lugar, tratamos algunas alternativas como soluciones legítimas, compararlas y aprender de las consecuencias resultantes a nivel sustantivo. Una cierta dosis de imprecisión forma parte del trabajo analítico. Sin embargo, dado que con AED es posible hacer mucho, en principio no hay nada en absoluto en contra de utilizar una combinación exploratoria creativa de otras medidas de distancia y métodos de aglomeración según convenga al caso y esté bien justificado. Sin embargo, este problema no se limita a los métodos de clasificación, sino que también se aplica a todo el campo de los modelos lineales. A menudo, en los libros de texto y en los artículos pertinentes sólo se encuentran pruebas anecdóticas de la elección de un determinado modelo estadístico. Pero ¿por qué exactamente el elegido es el más apropiado y superior a todos los demás, de modo que muchos otros posibles candidatos simplemente no entran en juego, queda sin respuesta en las normas y quizá sea incluso incontestable en la mayoría de los casos. Deberíamos aprender de esto a no limitarnos a adoptar los modelos de los demás, sino a considerar qué modelo se ajusta a nuestro caso concreto y aplicarlo de forma coherente.

La tabla 5.7 muestra una pequeña selección de posibles combinaciones para combinar con sensatez las distancias con los métodos de aglomeración con los que hemos tenido buena experiencia en la práctica. En cada caso concreto debe comprobarse si el procedimiento respectivo responde a las necesidades de la investigación. Para ello es necesario integrar la solución, el algoritmo subyacente y la pregunta de investigación en un contexto significativo, ¡todo un reto! Así pues, en contextos exploratorios, tendemos a aplicar varios procedimientos en paralelo y, a continuación, intentamos comprender las diferencias y similitudes de las soluciones sobre el trasfondo de los datos para aproximarnos a ellos.

Tabla 5.7: *Análisis jerárquico de conglomerados (método de formación de distancias y aglomeración)*

Combinación	Distancia*	Aglomeración*
1	manhattan	Ward
2	euclidean	single
3	euclidean	average
4	euclidean	complete

*en notación de R

Si se aplicara una lógica estadística clásica, se podría, por ejemplo, comprobar previamente la agrupabilidad jerárquica de los datos con la prueba de Oldenbürger (1981, p.199). Handl (2002) también enumera varios criterios de calidad (por ejemplo, la correlación cofenética, prueba de Mojena, 1977, el número de conglomerados y la pertenencia a clases, véase también Milligan & Cooper, 1985, el coeficiente gamma para determinar la calidad de una solución de conglomerado, propuesto por Oldenbürger (1981,

p.199). la bondad de una solución de conglomerados, propuesto por Hubert, 1974 y desarrollado por Goodman & Kruskal, 1954, véase también Bacher, 1994) y el código R asociado para evaluar las soluciones de conglomerados resultantes para evaluar su gravedad. Debido al carácter exploratorio del estudio, no entraremos aquí en estos detalles, sino nos concentraremos en la comparación cualitativa de casos entre las categorías, los rasgos portadores de rasgos y otra información contextual que se conozca sobre el caso.

El escalado multidimensional, por su parte, se hace sobre todo en 2D y 3D, porque sencillamente no tenemos imaginación para más de tres dimensiones. Además, queremos realizar una exploración de forma sencilla y clara. Se trata primero de la imagen global y no de detalles que posiblemente podrían investigarse mucho mejor en términos puramente cualitativos. Además, solemos enriquecer la MDS con un análisis de prototipos según Oldenbürger (Gürtler, 2005), en el que se pueden establecer conexiones significativas entre las unidades investigadas y marcarlas en el gráfico con una línea de conexión. Las conexiones significativas entre las unidades investigadas y marcarlas en el gráfico con una línea de conexión. De este modo, se crea una red de relaciones entre las unidades, "quién conecta con quién" y "quién no conecta con quién". La(s) unidad(es) investigada(s) con más conexiones recibe(n) el estatus de prototipo, lo que significa que globalmente tiene(n) la menor distancia a todos los demás portadores de rasgos en el contexto de las categorías investigadas (portadores de rasgos). El/los que no tienen o tienen sólo conexiones mínimas representan lo contrario del prototipo (véase también la Fig. 5.30), una especie de antiprototipo. Aparecen aislados o como valores atípicos. El método es muy sensible a los cambios en el contexto, es decir, en la base de datos. Por lo tanto, es ciertamente no apto para un procedimiento de confirmación, pero puede utilizarse como lupa para examinar determinadas áreas de datos de forma muy precisa dentro de su contexto. Técnicamente, el primer paso es utilizar `dist()`, se calcula una matriz de distancia euclidiana o de otro tipo para los datos de tamaño. Los datos pueden ser pre-escalados si es necesario, lo que los centra y estandariza la varianza a uno (varianza normalizada). Hay argumentos tanto a favor como en contra, pero aquí prescindimos de ello.

Empezamos leyendo los datos y observando su estructura. Como los nombres de las dimensiones son bastante largos y esto no nos gusta (legibilidad), tomamos sólo las cuatro primeras letras de cada dimensión y creamos otra columna para las abreviaturas muy cortas que resultan útiles en los gráficos (`ptII_quan_EDA_caso_España_liderazgo_en_la_educación.r`).

```
# read data
rawd <- read.table("Spain_leadership-in-education_data.tab", header=TRUE, sep="\t")
rawd <- t(rawd) # transpose for later analysis
rawd
str(rawd)
dim(rawd)
# cols = cases/ interviews
# rows = dimensions content leadership
# abbreviate dimension names
dnams <- rownames(rawd)
dnams.tab.abbrev <- data.frame(abbrev=paste("D",1:length(dnams),sep=""),
           dnams.short=substr(dnams, 1, 4),
           dnams.abbrev=abbreviate(dnams, named=FALSE), categories=dnams)
dnams.tab.abbrev
rownames(rawd) <- dnams.tab.abbrev[,"dnams.short"]
```

Ahora podemos calcular las distancias:

```
rawd.d <- dist(rawd, method="euclidean")
```

Si desea escalar los datos de antemano, se podría aplicar el código R siguiente:

```
# scale before calculating distance matrix R-Code
scaling <- FALSE
if(scaling) rawd.scaled <- sweep(rawd,2,sqrt(apply(rawd,2,var)),"/")
```

R normalmente sólo produce una matriz triangular. Ésta puede transformarse en una matriz completa con `distfull()` (Handl, 2002),

```
# distance matrix R-Code
rawd.d <- dist(rawd, method="euclidean")
rawd.d
rawd.d.full <- distfull(rawd.d)
print(rawd.d.full,digit=3)
```

que tiene este aspecto:

```
       char emot anti  prof part cult  edut  admi
char    0.0  21.7 14.73 43.2 30.2 16.34 15.03 17.49
emot    21.7  0.0 21.31 35.2 27.2 27.20 28.05 25.79
anti    14.7 21.3  0.00 43.7 31.6 11.49 11.96  9.85
prof    43.2 35.2 43.67  0.0 22.0 48.98 48.06 48.39
part    30.2 27.2 31.59 22.0  0.0 35.86 35.40 37.00
cult    16.3 27.2 11.49 49.0 35.9  0.00  9.54  9.22
edut    15.0 28.1 11.96 48.1 35.4  9.54  0.00  9.70
admi    17.5 25.8  9.85 48.4 37.0  9.22  9.70  0.00
```

El *corte óptimo a través de una matriz de proximidad* se aplica a la matriz de distancia (Oldenbürger, 1981, p.155; Gürtler, 2005). Se genera una matriz triangular (0; 1) para cada valor de la matriz de distancias. Se crea tomando cualquier valor de distancia y dando a todos los demás valores mayores que este valor de distancia un valor de uno, mientras que a todos los valores menores que el valor de distancia reciben un cero. El valor de distancia correspondiente puede recibir un uno o un cero. Como esto se hace constantemente para cada valor, no cambia el número relativo de conexiones entre las categorías. El resultado es una matriz triangular binaria (0; 1). Esta matriz triangular (0; 1) se correlaciona con la matriz triangular de distancia original. El procedimiento se lleva a cabo sistemáticamente para cada valor de distancia empírica, de modo que cada valor de distancia actúa como un punto de corte. Hay tantos coeficientes de correlación como valores de distancia. En R lo hacemos con `optcut()`. El resultado es una tabla con los valores de corte *sp* y los correspondientes coeficientes de correlación *cc*.

```
# opt
vektorOpt <- optcut(rawd.d.full)
str(vektorOpt)
```

De este conjunto de valores de correlación se selecciona como corte la correlación máxima que marca la máxima potencia de mapeo de la matriz de distancia a la (0; 1)-matriz . Para ello, se ordena la tabla o se extrae directamente el máximo (salida abreviada a continuación).

```
> # max correlation
> vektorOpt[with(vektorOpt, order(-cc)),]
sp cc
14 25.787594 0.8539131
17 28.053520 0.8535729
...
27 48.394215 0.3434728
1 9.219544 0.2584096
> vektorOpt[vektorOpt[,"cc"] == (max(vektorOpt[,"cc"])),]
sp cc
14 25.78759 0.8539131
```

Con `plot.optcut()`se puede mostrar las correlaciones:

```
plot.optcut(vektorOpt=vektorOpt, outZO=outZO,
            TITLE="Leadership (8 dimensions)")
```

La matriz triangular (0; 1) asociada se transforma en una matriz completa y se calcula la suma de columnas (o suma de filas, no importa, ya que se trata de una matriz simétrica). La función de R `outZO()` genera todos los valores necesarios.

```
> # create prototype matrix (0,1) based on optimal cut
> outZ0 <- makeZ0(dm.full=rawd.d.full, vektorOpt=vektorOpt)
> outZ0
$prototype.mat
      char emot anti prof part cult edut admi
char     1    1    1    0    0    1    1    1
emot     1    1    1    0    0    0    0    1
anti     1    1    1    0    0    1    1    1
prof     0    0    0    1    1    0    0    0
part     0    0    0    1    1    0    0    0
cult     1    0    1    0    0    1    1    1
edut     1    0    1    0    0    1    1    1
admi     1    1    1    0    0    1    1    1
$maxcc
[1] 0.8539131
$cutoff
[1] 25.78759
$protovec
      char anti admi cult edut emot prof part
       6    6    6    5    5    4    2    2
```

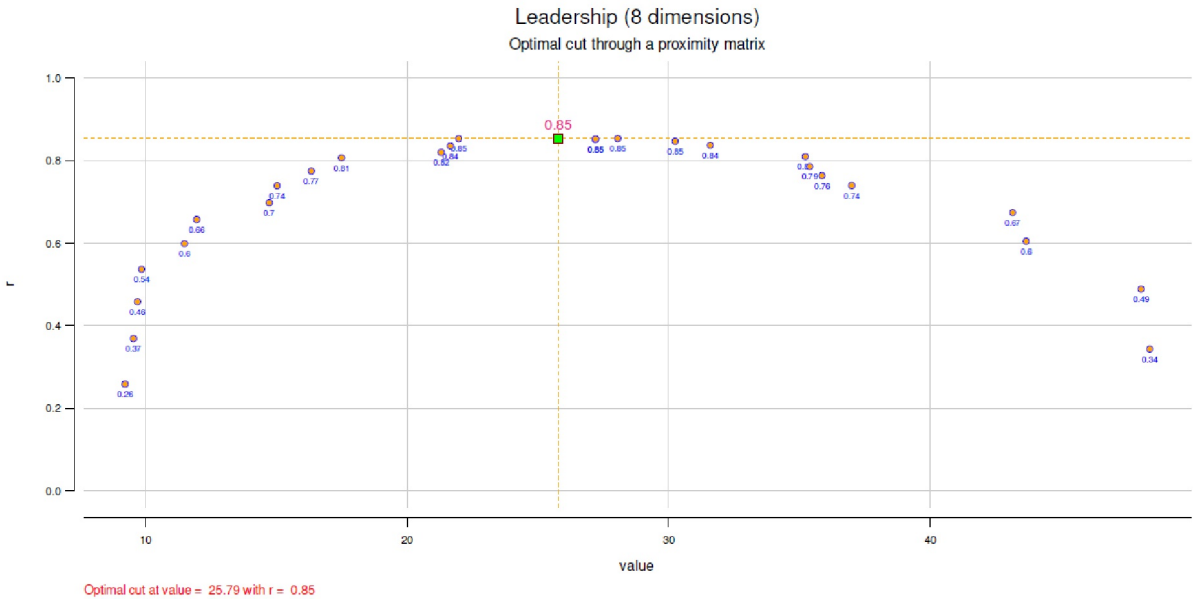


Figura 5.29. *Estudio de Gento et al. (2015b, corte óptimo a través de una matriz de proximidad).*

Estas sumas dan como resultado un valor de prototipicidad para cada columna (fila), es decir, con cuántas otras columnas (filas) está conectada o no la categoría respectiva (columna, fila). La(s) categoría(s) con la suma máxima forma(n) el(los) prototipo(s). A la inversa nos referimos a la(s) categoría(s) con la suma mínima como valor(es) atípico(s), valor(es) extremo(s) o simplemente como antiprototipo(s). El vector resumen de prototipicidad, que suma el número de conexiones a través de los portadores de características individuales (aquí: dimensiones de liderazgo), está contenido directamente en el objeto, y así es como se puede identificar el prototipo:

```
# prototipo
outZ0$protovec
```

Evidentemente, hay más de un prototipo: carisma, anticipación y administración. Profesionalidad y participación se sitúan en el otro extremo del espectro. A partir de ahí, se pueden utilizar visualizaciones gráficas como el escalado multidimensional (MDS; 2D, 3D) para establecer las conexiones entre las categorías correspondientes y codificar por colores los prototipos (véase la Fig. 5.30 arriba para 2D y abajo para 3D). Esto puede hacerse para el MDS bidimensional con `plot.prototype2D()` y permite una imagen

clara de la proximidad y la distancia, así como de las conexiones existentes entre las categorías. Primero, sin embargo, es necesario crear el MDS para 2D y 3D:

```
# calculate MDS 2D/ 3D
rawd.d2 <- cmdscale(rawd.d, k=2, eig=TRUE, add=TRUE, x.ret=TRUE)
rawd.d3 <- cmdscale(rawd.d, k=3, eig=TRUE, add=TRUE, x.ret=TRUE)
```

Ahora viene la salida gráfica. Esto es posible de forma rudimentaria con `pairs()`, pero a esta función R carece de añadidos como el análisis de prototipos. Para el MDS tridimensional hay `plot.prototype3D()`:

```
# plot MDS with prototypes
plot.prototype2D(rawd.d2, outZO=outZO,
                  TITLE="Leadership (8 dimensions)", fac=1.6)
plot.prototype3D(rawd.d3, outZO, TITLE="Leadership (8 dimensions)",
                  ANGLE=26, labelmds=rownames(rawd), box=FALSE)
```

Además, merece la pena echar un vistazo a los valores propios del MDS con `plot.eig.mds()` (véase la Fig. 5.31):

```
# plot eigenvalues
plot.eig.mds(sdata=rawd.d2, TITLE="Leadership (8 dimensions)",
           SUB="Eigenvalues (Multimensional scaling)")
```

El procedimiento esbozado es excelentemente adecuado para inspirar reflexiones sobre por qué estas o aquellas unidades de investigación están conectadas entre sí, pero no con otras, y cómo se distribuye el número total de conexiones (palabra clave: red de relaciones). En tales análisis, resulta especialmente llamativo (Gürtler & Huber, 2006) que en el análisis de datos cualitativos, curiosamente, se utilicen términos cuantificadores como "más" o "menos", "con frecuencia" o "rara vez", etc., aunque en realidad debería tratarse de un análisis cualitativo y no cuantitativo. En el análisis de prototipos se busca a propósito la cantidad de conexiones cualitativas entre portadores de características o categorías. La integración de métodos tiene lugar en muchos niveles sin que siempre se mencione explícitamente y quizá incluso de forma inconsciente. No nos excluimos de ello, sino que simplemente intentamos utilizarlo de forma más consciente.

El análisis MDS y de prototipos va seguido de la formación de clusters jerárquicos. Los análisis de conglomerados jerárquicos en R comienzan, como el MDS, con la formación de una matriz de distancias mediante `dist()`. Además de la distancia euclídea, esta función de R también ofrece las opciones maximum, `manhattan, canberra, binary` y `minkowski`. En la publicación de Handl (2002) existe un script en R para la distancia `Mahalanobis`, que tiene en cuenta las covarianzas existentes entre las categorías. La aglomeración, a su vez, puede realizarse en R mediante `hclust()`. Por el lado de los métodos de partición para determinar el número óptimo de conglomerados existe la función `kmeans()`. Una introducción breve y concisa a los análisis de conglomerados con R y los algoritmos subyacentes es proporcionada por Boehmke (2019). Los dendrogramas resultantes pueden representarse gráficamente – entre otros con `plot.hclust()` o con una salida gráfica más especializada como el del paquete `ape` de R.

Además, hay muchos otros casos especiales, que se resumen en la página de resumen de R (CRAN, 2019b) sobre análisis de conglomerados. Nota que no hay diferencia si para un objeto creado con `hclust()` se utiliza `plot()` o `stats:::plot.hclust()`, que reside en el área del paquete `stats`. Usar simplemente `plot()` es más fácil, ya que R se asegura internamente de que se produce la salida correcta según la clasificación de un objeto siempre que se disponga de una función para ello.

Para el presente conjunto de datos del estudio español, es como sigue. En primer lugar creamos una matriz para las combinaciones de procedimientos de formación de distancia y aglomeración. Este sólo prepara la automatización de los análisis siguientes.

```
dists <- c("manhattan","euclidean","euclidean","euclidean")
agglos <- c("ward","single","average","complete")
methoden <- data.frame(dists, agglos)
d.methoden <- dim(methoden)
```

```
methoden
```

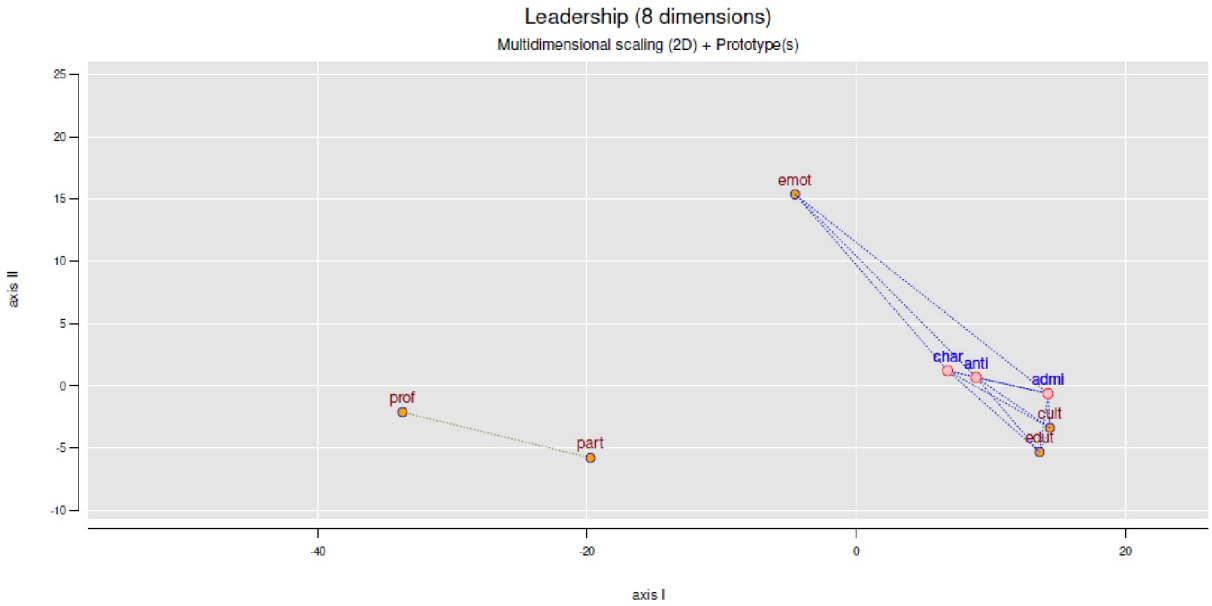


Figura 5.30. *Estudio de Gento et al. (2015b, escalado multidimensional 2D y 3D).*

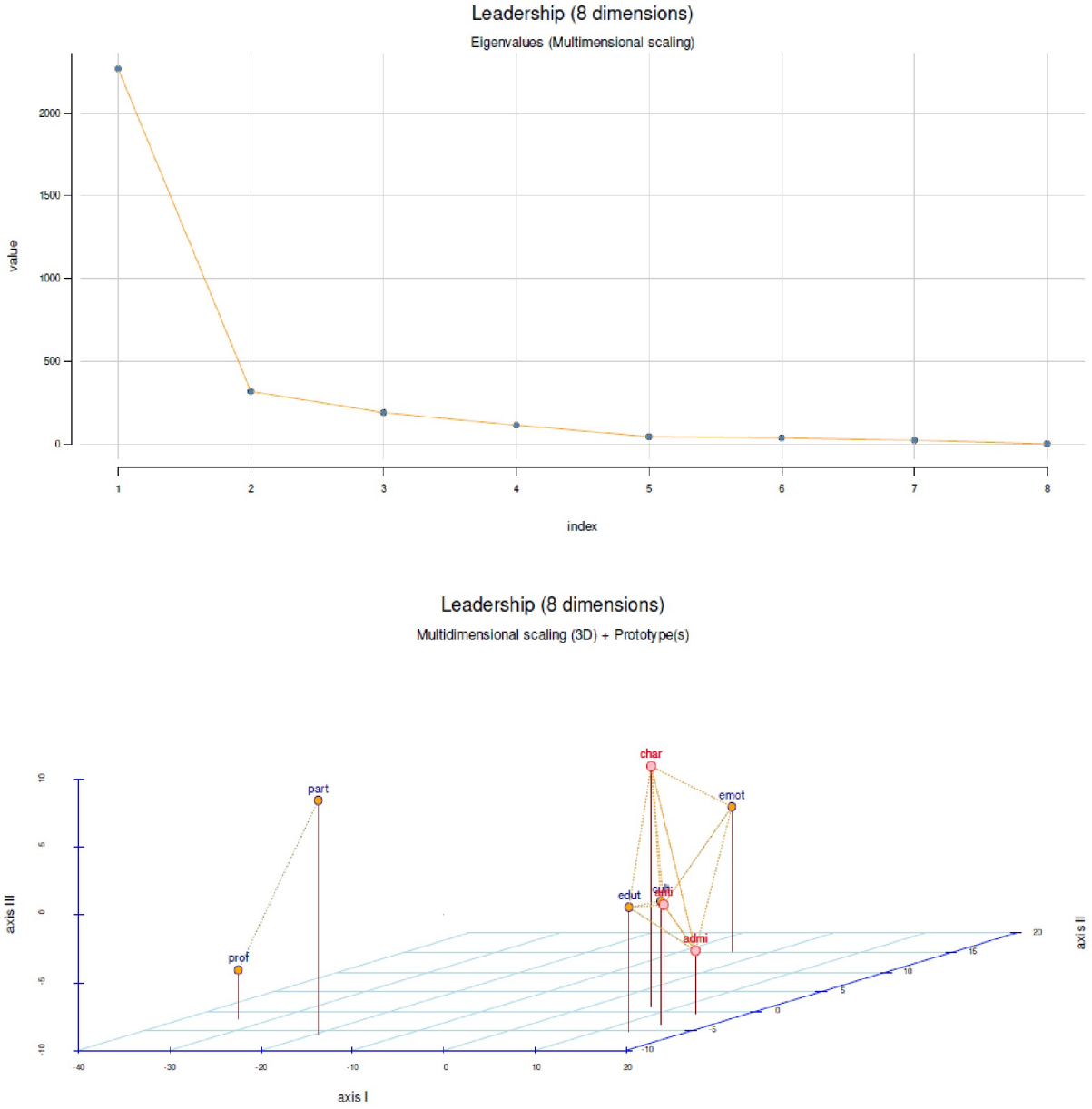


Figura 5.31. *Estudio de Gento et al. (2015b, escala multidimensional, Eigenwerte)*

Un escalado necesario podría hacerse de antemano con `rawd.s <- scale(rawd)`. A continuación la formación respectiva de las matrices de distancia sigue

```
# scaling if required
rawd.s <- scale(rawd)
# distance matrices
dists.res <- lapply(seq_along(1:d.methoden[1]),
               function(i) dist(rawd, method=methoden[i,"dists"]))
dists.res.full <- lapply(dists.res, distfull)
names(dists.res.full) <- methoden$dists
dists.res.full
```

y la aplicación de procederes de aglomeración:

```
# agglomeration
clusts.res <- lapply(seq_along(1:d.methoden[1]),
function(i) hclust(dists.res[[i]],
           method=methoden[i,"agglos"]))
comb.nam <- paste(methoden[,1],methoden[,2],sep=" / ")
names(clusts.res) <- comb.nam
clusts.res
```

Por último, trazamos todos los dendrogramas (véase la Fig. 5.32) para ver los resultados:

```
# plot dendrograms R-Code
par(mar=c(5,5,4,2), oma=c(2,1,1,1), "cex.axis"=0.8, mfrow=c(2,2))
for(i in 1:d.methoden[1])
{
    plot(clusts.res[[i]], main="", axes=TRUE, sub="",
           xlab=comb.nam[i], ylab="height", col="violetred2")
    rect.hclust(clusts.res[[i]], k=3, border="green")
}
mtext("Leadership (8 dimensions)", 3, line=-1.5, cex=1.5, outer=TRUE)
mtext("Hierarchical Cluster Analysis", 3, line=-3, cex=1.1, outer=TRUE)
```

Aunque el trabajo basado en vectores suele ser mucho más rápido (Ligges, 2005), es aconsejable utilizar bucles con for(...) {...} para la salida no crítica en tiempo, ya que son más fáciles de escribir y leer especialmente para los principiantes. Ahora comprobamos si la observación anterior sobre `plot()` es correcta (no se imprime la salida gráfica):

```
# check whether plot() and stats:::plot.hclust() are the same
par(mar=c(5,5,4,2), oma=c(2,1,1,1), "cex.axis"=0.8, mfrow=c(1,2))
plot(clusts.res[[1]], col="darkred")
stats:::plot.hclust(clusts.res[[1]], col="steelblue")
attr(clusts.res[[1]],"class")
```

Las parcelas creadas de este modo deben ser absolutamente idénticas. La última llamada da la clase del objeto, que en este caso es `hclust` sin más sorpresa. Es posible obtener más información sobre un objeto de R con `str()` – que puede ser muy extensa para objetos grandes (por ejemplo, modelos lineales). Dado que R es un lenguaje orientado a objetos, se puede acceder directamente a cualquier área de estos

```
str(clusts.res)
str(clusts.res[[1]])
clusts.res[[1]][["method"]]
```

Además, si nos interesa el código R de una función, basta con llamarla. Si no se puede descubrir porque está oculta, necesitamos el llamado `namespace`, es decir, el área en la que se encuentra la función (objeto) en R:

```
hclust
stats:::hclust
```

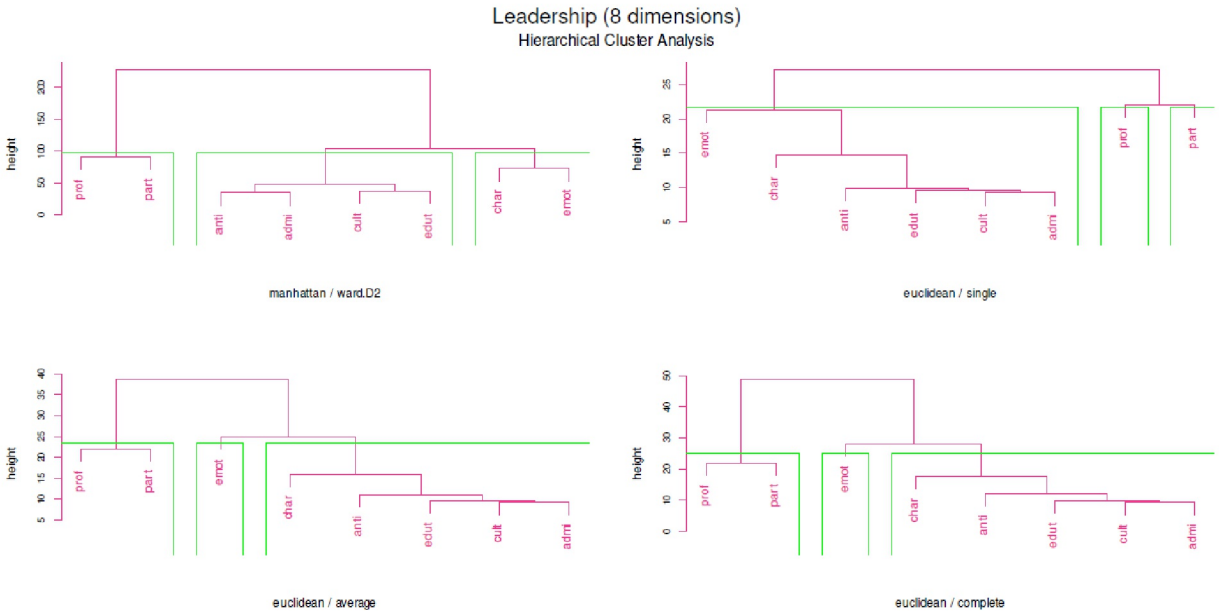


Figura 5.32. *Estudio Gento et al. (2015b, Análisis jerárquico de clusteres)*

Si las funciones tienen nombres muy similares porque tienen una llamada común, como en el ejemplo de `plot()`, ayudará `methods()`:

```
methods("lm")
methods("plot")
```

Las funciones ocultas tienen un "*" en la llamada anterior y pueden encontrarse utilizando la combinación de `namespace` y nombre de función. Es aún más fácil para las funciones S3 `()` con `getAnywhere()`.

```
# Funciones S3 get
getAnywhere(plot.hclust)
```

El código R no se muestra si se trata de código compilado, como código Fortran o C# optimizado que sólo está disponible compilado en lenguaje máquina. Si hay interés en este código, se necesitan las fuentes en R de los paquetes correspondientes con los que la respectiva función fue compilada. También puede ser de interés para comprender mejor las diferencias entre las funciones S4 y S3. Estas cuestiones se discuten en los foros y listas de correo.

La figura 5.32 muestra los resultados del HCA para las ocho dimensiones. La figura 5.30 muestra los resultados del MDS en 2D y 3D. Desde nuestro punto de vista, es importante dibujar estos gráficos tanto para las codificaciones de contenido (complejos temáticos) como en paralelo para las entrevistas (casos) e intentar integrar los resultados – dialécticamente. Consideramos negligente considerar sólo una de estas perspectivas, ya que los resultados de un análisis dependen directamente de los del otro y de sus cambios inherentes. En el caso concreto, esto significaba examinar las conexiones y diferencias entre los ocho complejos temáticos y preguntarse cómo pueden ordenarse los entrevistados en términos de cercanía y distancia, qué (sub)agrupaciones se forman y cómo pueden agruparse, qué personas se localizaron por separado e individualmente (valores atípicos, etc.). Técnicamente, la matriz inicial sólo tiene que transponerse con `t()` antes de la primera formación de distancias. El resto del código R puede asumirse directamente. Ambos análisis, relacionados entre sí, responden a qué temas desempeñan un papel en qué subgrupos y qué temas no están sujetos a tal diferenciación. Básicamente siempre se trata de *similitudes frente a diferencias* en las estructuras de datos subyacentes, ya sean entrevistas (personas, casos), codificación (complejos temáticos)

o códigos abstraídos y agregados (metacódigos). Dado que entonces es necesario volver a los datos originales con los resultados, queda claro hasta qué punto el procedimiento analítico utilizado ha tenido un efecto de conocimiento. Esto requiere un estudio en profundidad del procedimiento y del algoritmo subyacente para entender por qué, por ejemplo, determinados subgrupos aparecen con este procedimiento de aglomeración y menos con otro y cómo puede interpretarse esto. A la inversa, incluso con procesos de análisis puramente cualitativos, siempre nos preguntamos por los *elementos comunes* y las *diferencias* y por los *procesos temporales* si éstos desempeñan un papel en el contenido. Éstas son las cuestiones principales que siempre surgen en contextos muy diferentes.

También es esencial volver a los datos originales con los resultados del AED. Allí es importante aplicar los nuevos conocimientos adquiridos, integrarlos y, sobre todo, comprobar si parecen tener sentido. En contra de la lógica cuantitativa generalizada, no se trata de que el procedimiento no decide si tiene sentido en este punto. Por ello, en el presente estudio de caso se investigó si las conexiones entre los complejos temáticos tienen una base realista y justificable. La cuestión del sentido debe abordarse lo más directamente posible con los datos originales (a nivel de texto, audio, vídeo o imagen) y no sólo a nivel de codificaciones o metacódigos abstractos. No hay que escatimar esfuerzos a la hora de pensar, porque esto promete nuevas perspectivas y merece la pena. No se trata de encontrar algo lo más rápidamente posible, sino de encontrar lo que realmente hace avanzar la respuesta a la pregunta de investigación.

Como puede verse en la figura 5.30 (MDS, 2D, complejos temáticos), en el empírico ejemplo las dimensiones de liderazgo podrían dividirse en tres grupos:

1. G1 – D4 profesional, D5 contributivo.
2. G2 – D1 carismático, D2 emocional
3. G3 – D3 prospectivo, D6 cultural, D7 formativo, D8 administrativo.

Esto refleja la solución de cluster (véase la Fig. 5.32) de combinar la distancia Manhattan y el algoritmo Ward. Sin embargo, las demás soluciones de agrupación son relativamente parecidas y no difieren sustancialmente, sino que apuntan en una dirección similar. G1 estaba situado bastante separado de G2 y G3 y sólo estaba conectado internamente. G2, en cambio, estaba estrechamente asociada (muchas conexiones prototipo), pero espacialmente se distinguía claramente de ella. En general, el estudio intensivo de los datos originales, la codificación y el análisis gráfico dieron la impresión de que los datos "hablan por sí solos y revelan sus estructuras". En nuestro caso, se trataba de los tres complejos enumerados anteriormente. Si trasladamos los resultados a un nivel de contenido, saltan a la vista dos resultados en particular:

- G1 mostró la proximidad de la profesionalidad y la participación en el contexto de las estructuras de liderazgo. Esto contradecía las estructuras jerárquicas que se han practicado en contextos educativos de todo el mundo durante generaciones. Al parecer, no se apreciaba realmente la jerarquía estricta. Más bien, en los contextos educativos, se esperaba que un líder permitiera la participación de los demás y, por lo tanto, permitiera o concediera explícitamente que los demás pudieran compartir la responsabilidad, hacia un objetivo común con intereses compartidos. Esto abogaba por una comprensión más democrática del liderazgo, que, sin embargo, no es idéntica a la disolución de estructuras o incluso a la arbitrariedad. Por supuesto, en la práctica siempre habría que ver cómo la participación y la participación pueden aplicarse y coordinarse. El trabajo no ha hecho más que empezar aquí, pero pone de relieve un claramente esbozado punto de partida.
- Lo mismo se aplica a la estrecha proximidad del carisma y las emociones (G2). En el estudio de Gento (2014) de hecho, las submuestras de determinados centros educativos mostraban en general puntuaciones elevadas en estas dos dimensiones. Del mismo modo, se observaron diferencias entre la importancia atribuida y la evidencia o presencia de estas características en los líderes. Especialmente importante resultó el caso de alta evidencia/presencia y baja importancia para las características emocionales de los líderes. Esto aportó pruebas de que el equilibrio entre carisma y comportamiento emocional no siempre está presente en los líderes y que es exactamente así como se percibe desde el exterior. Esto puede tener implicaciones prácticas, como qué se espera de los líderes y cómo se les puede formar o tutelar en el trabajo. Como otra visualización multivariante del AED en el caso que nos ocupa, se puede utilizar un mapeo por colores, ordenado por frecuencias.

Aquí, las frecuencias se subdividen bidimensionalmente a lo largo de una supuesta escala de temperatura y se codifican por colores en consecuencia para poder encontrar fácil e intuitivamente puntos distintivos en la matriz. Todo ello se engloba bajo el término heatmap (imagen térmica). En algunos casos, se puede trazar un HCA en los laterales, lo que hace que estas visualizaciones sean extremadamente versátiles y potentes.

```
hc <- hclust(dist(rawd, method="manhattan"),method="ward.D2")
hr <- hclust(dist(t(rawd), method="manhattan"),method="ward.D2")
heatmap3(t(rawd), Rowv=as.dendrogram(hr), Colv=as.dendrogram(hc),
           scale="row", balanceColor=TRUE, showRowDendro=TRUE, ColSideCut=50)
```

La Figura 5.33 muestra esto para el presente conjunto de datos (códigos) sin discutirlo más aquí. Se pueden realizar otras variantes con `heatmap()` o `heatmap.plus()` del paquete R `heatmap.plus` o con `levelplot()` del paquete R `lattice`. Una extensión interesante la ofrecen las rampas de color continuas con `colorRampPalette()` del paquete `RColorBrewer`. Ampliando esto, las matrices de dispersión y los correlogramas pueden pueden visualizar las relaciones entre las variables y sus distribuciones en un paquete compacto.

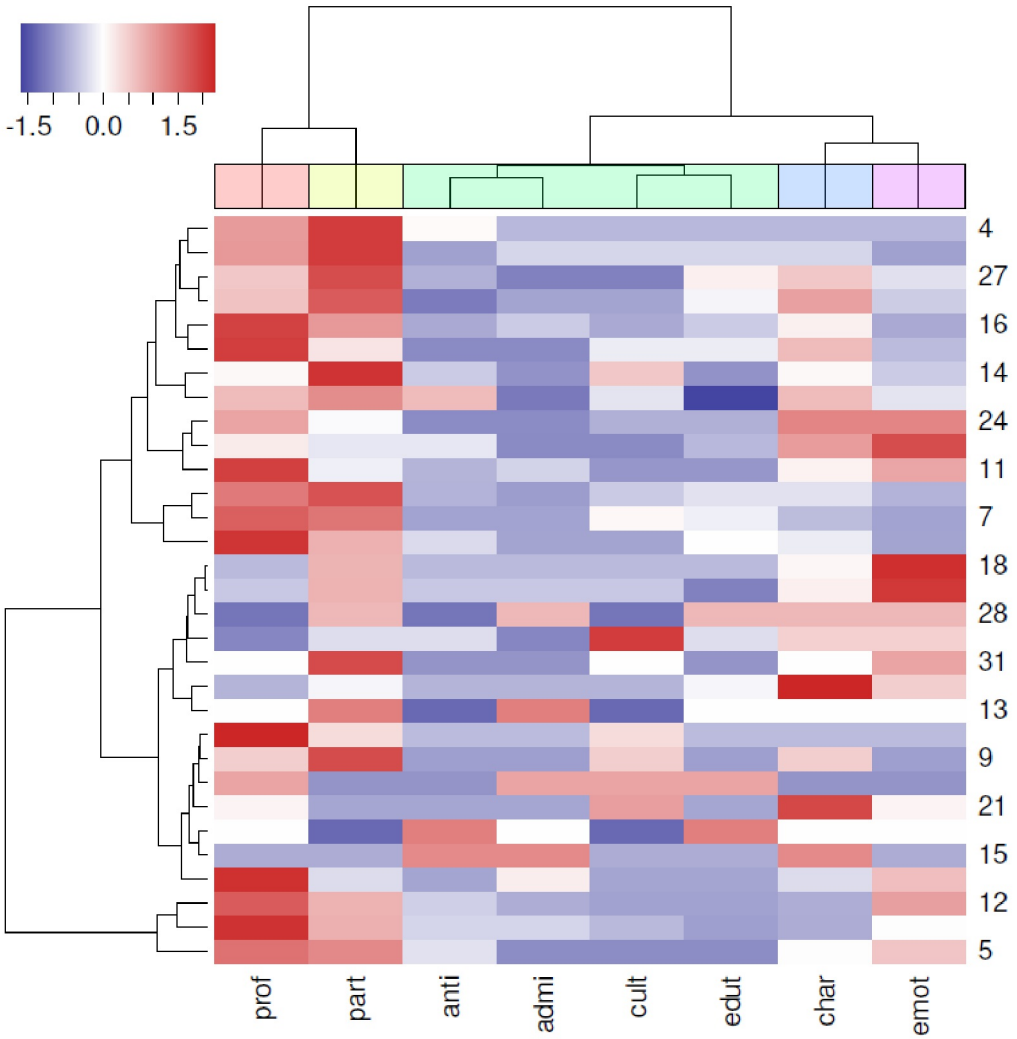


Figura 5.33. *Estudio de Gento et al. (2015b, heatmap)*

```
coul <- colorRampPalette(brewer.pal(8, "PiYG"))(25)
heatmap(rawd, xlab="", ylab="", main="Leadership", col=coul)
```

Con estos resultados, el estudio fue capaz de proporcionar las traducciones deseadas a la práctica y, equipado con nuevos conocimientos, fueron posibles otras acciones (por ejemplo, política educativa, a nivel de coaching, educación y formación, etc.). La ventaja del AED en este caso fue el descubrimiento de estructuras y patrones.

5.5.6 Un experimento sobre la variabilidad del ritmo cardíaco

Lo que sigue no es un estudio puramente exploratorio de datos, sino la aplicación de técnicas de AED en un diseño principalmente de prueba. Sin embargo, se transformó en un diseño "realmente" exploratorio debido a los resultados del AED, ya que salieron a la luz nuevos hallazgos que no eran coherentes con una lógica estrictamente confirmatoria y aplicada a ciegas. En este caso, parecía tener sentido omitir la significación y dilucidar lo esencial en los datos. Tal mezcla – ni puramente hipotético-deductiva ni puramente exploratoria – está ciertamente descrita en la literatura. Si se abordan los datos de forma flexible con análisis gráficos, pueden surgir resultados que ciertamente no se consiguen mediante la aplicación rigurosa de estructuras analíticas rígidas con lógica de significación binaria.

Un estudio experimental de la quiropráctica realizado por Wipfler[1] (2017) examinó la influencia de una técnica clave de tratamiento quiropráctico, un *ajuste de altas cervicales con alta velocidad y baja amplitud*, en su potencial para iniciar cambios a nivel de la *variabilidad de la frecuencia cardíaca* (VFC). La VFC actúa como indicador del estrés. Lamentablemente, este conjunto de datos no puede estar disponible en línea. No obstante, imprimimos el código R. En primer lugar, se lee todo el entorno del análisis previamente almacenado (`ptII_quan_EDA_case_Chiro_heartrate-variability.r`).

```
# cargar los datos a través del entorno almacenado
load("HW_FINAL_status-quo-161216_complete-environment.RData")
```

El diseño pre-post equilibrado se llevó a cabo con n = 18 personas (n = 9 por cada grupo de tratamiento y control). Un vistazo al diseño muestra que está equilibrado.

```
> tabla(gd)
gd
Control.t1.pre Control.t2.post Tratamiento.t1.pre Tratamiento.t2.post
             9              9                 9                 9
```

Desafortunadamente, ni la literatura relevante ni los fabricantes de los respectivos instrumentos son claros sobre qué marcador(es) son máximamente informativos para la VFC y los posibles cambios debidos a las intervenciones quiroprácticas. Los fabricantes de los instrumentos de medición publican muchos datos, algunos de los cuales están muy correlacionados entre sí, lo cual no es sorprendente. Además, según el autor, la calidad de los instrumentos de medición no está garantizada de manera uniforme. Hay aparatos más caros (por ejemplo, uso de electrodos de alta calidad) y más baratos, por lo que la calidad de los datos así generados puede estar sujeta a fuertes fluctuaciones. En vista del pequeño tamaño de la muestra, se seleccionaron tres variables (SDNNms, RMSSDms, SD2/SD1) como variables dependientes desde el punto de vista teórico-contenido de Wipfler para evitar resultados aleatoriamente significativos (palabra clave: más variables dependientes que sujetos de estudio). Éstas se correlacionaban entre sí a distintos niveles, de modo que se garantizaba que la misma variable no aparece tres veces (véase la tabla 5.8). A continuación se seleccionan las variables pertinentes y se revisan los datos brutos.

[1] Agradecemos encarecidamente al Sr. Holger Wipfler, de *Vital im Puls*, su permiso para presentar aquí los datos, incluidos los resultados, los gráficos, la información contextual e interpretaciones.

```
# select vars
relvars <- c("SDNNms","RMSSDms","SD2durchSD1")
daten.analy <- data.frame(daten.red[,relvars], log(daten.red[,"SD2durchSD1"]))
cnams <- c("SDNNms","RMSSDms","sd2/sd1","log(sd2/sd1)")
colnames(daten.analy) <- cnams
rownames(daten.analy) <- paste(as.character(gd), rownames(daten.analy),sep=".")
daten.analy
```

Siguen las intercorrelaciones, pero omitimos los valores *p*.

```
> # correlations
> cor.red.tab <- corp(daten.analy)
> # correlations: -1 < r < +1
> print(cor.red.tab$r, digits=2)
             SDNNms RMSSDms sd2/sd1 log(sd2/sd1)
SDNNms         1.00    0.86   -0.27      -0.23
RMSSDms        0.86    1.00   -0.65      -0.66
sd2/sd1       -0.27   -0.65    1.00       0.98
log(sd2/sd1)  -0.23   -0.66    0.98       1.00
> printcorp(cor.red.tab$p)
            SDNNms RMSSDms   sd2/sd1 log(sd2/sd1)
SDNNms          NA   0.000     0.321       0.386
RMSSDms      0.000      NA     0.007       0.006
sd2/sd1      0.321   0.007        NA       0.000
log(sd2/sd1) 0.386   0.006     0.000          NA
```

Tabla 5.8: *Estudio de Wipfler (2017, intercorrelaciones de las variables dependientes)*

	SDNNms	RMSSDms	SD2/SD1	log(SD2/SD1)
SDNNms				
RMSSDms	0.860			
SD2/SD1	−0.265	−0.647		
log(SD2/SD1)	0.232	−0.656	0.979	

La figura 5.34 muestra un correlograma de las aVs, donde en la diagonal se estima la densidad de las variables y las correlaciones se visualizan como una elipse (con línea de regresión local) y como un gráfico circular en las zonas inferior y superior.

```
# variables of interest
corrgram(daten.analy, lower.panel="panel.ellipse",
         upper.panel="panel.pie", diag.panel="panel.density",
         col.regions=colorRampPalette(pal),
         main="Corrgram", order=FALSE)
```

El diseño experimental permitía formular hipótesis claras *a priori* sobre los efectos de interés. Así pues, en principio era hipotético-deductivo. La hipótesis rectora era que la intervención quiropráctica debía mejorar todos los valores para el grupo de tratamiento, pero no para el grupo de control. Técnicamente, esto significaba un aumento de los valores *después* de la intervención.

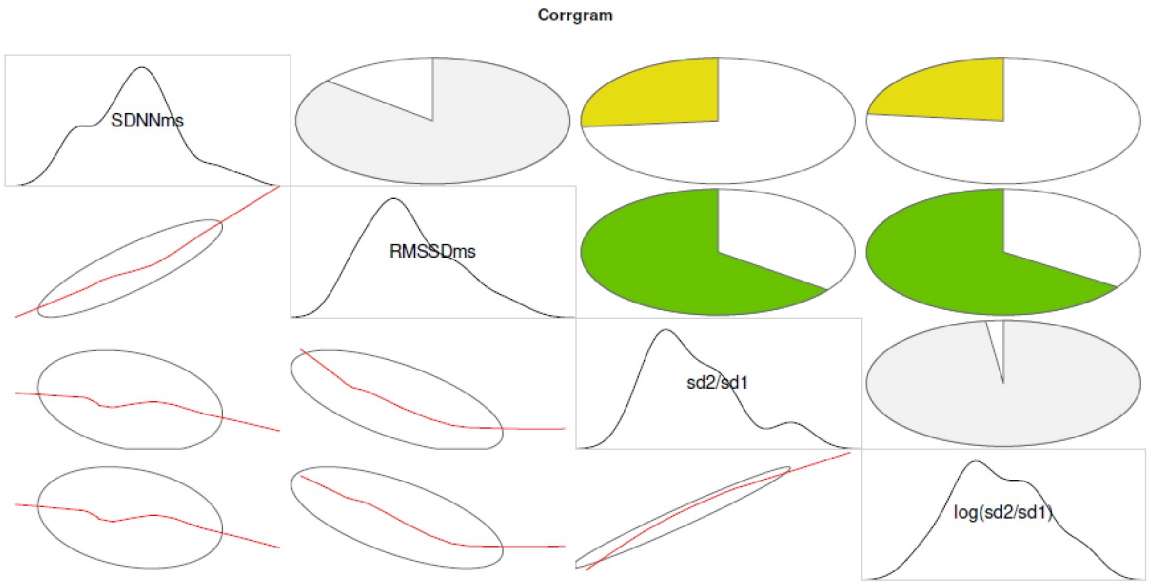


Fig. 5.34. *Estudio de Wipfler (2017, correlograma de variables dependientes)*

Los datos se analizaron estadísticamente de forma clásica con la ayuda de modelos lineales jerárquicos (Pinheiro & Bates, 2009) utilizando `lmer` del paquete R `lme4` con R. Debido a la pequeña cantidad de datos, sólo se pudo permitir un intercepto aleatorio por persona (sujeto). Habría sido más conveniente permitir que variaran no sólo el intercepto sino también la pendiente. Esta podría ser la tarea de un estudio de seguimiento o replicación más amplio. Con una serie de mediciones más larga, con un número significativamente mayor de puntos de medición, habría sido posible modelizar no sólo un nivel diferente por persona, sino también un curso diferente. Esto tendría en cuenta el hecho de que cada persona reacciona de forma diferente a las intervenciones a nivel corporal en función de sus propios antecedentes individuales y que el cuerpo se asienta en un nuevo nivel a más largo plazo. La situación dio lugar al siguiente modelo lineal básico

```
(R-notación según lmer()):
R-notación aV ~ tratamiento * tiempo + edad + (1|sujeto).
```

Es importante comprender que en un diseño de este tipo los grupos (tratamiento frente a control) no deben diferir en el tiempo Pre. Un efecto consistente del tratamiento en ambos puntos temporales (pre vs. post) hablaría, según la lógica del diseño practicado, de una falta de aleatorización o, simplificado, de diferencias sistemáticas en el tiempo t1 antes de la intervención. En consecuencia, las diferencias en t2 después del experimento (post) no serían claramente atribuibles a la intervención si ya existían parcialmente antes de la intervención.

Sin embargo, debido al tratamiento, era de esperar una evolución diferente según el grupo. Esto hablaba de un efecto de interacción entre el tratamiento y el tiempo. No era de esperar un efecto principal tiempo, ya que el tiempo sólo puede entenderse en función del tratamiento, y ése es el efecto de interacción. Un efecto principal *tiempo* significaría que ambos grupos cambiarían sustancialmente con el tiempo, y esto contradiría el efecto del tratamiento, es decir, si el grupo de control cambiara sistemáticamente a pesar de no haber intervención. Esto requeriría una teoría, de la que no se disponía. Para el grupo de control – sin tratamiento – no debería haber cambios sistemáticos a lo largo del tiempo. Por lo tanto, era de esperar que ambos grupos tuvieran aproximadamente los mismos valores en el momento t1, lo que, sin embargo, puede quedar enmascarado por efectos aleatorios en grupos pequeños. A continuación, el curso del tratamiento debería ir en la dirección prevista por la hipótesis en t2 y el del control debería variar, en el mejor de los casos, de forma aleatoria, pero sin mostrar cambios sistemáticos – es decir, sin aumento de los valores – como en el caso del tratamiento.

En principio, existía la posibilidad de que los grupos ya difirieran por azar al principio. La muestra se reclutó entre los empleados de un concesionario de automóviles, lo cual es una muestra selectiva y no representativa para la quiropráctica y, en última instancia, refleja razones logísticas independientes del estudio. Lamentablemente, esto tuvo un efecto desfavorable, como puede verse abajo. Sin embargo, dentro de esta muestra, la asignación a las condiciones (tratamiento sí/no) fue aleatoria para minimizar los problemas de selección.

En el nivel *d de Cohen* (véase la tabla 5.9), surgieron las siguientes diferencias de pre (t1) a post (t2) para el tratamiento frente al control. Se nombran las comparaciones pertinentes (tratamiento frente a control a pre y tratamiento de pre a post).

Tabla 5.9: *Estudio de Wipfler (2017, d de Cohen)*

d de Cohen	Comparación	Variables			
		SDNNms	RMSSDms	SD_2/SD_1	$\log(SD_2/SD_1)$
$K_{t2} - K_{t1}$	sin intervención	−0.046	−0.008	0.240	0.157
$T_{t1} - K_{t1}$	diferencias antes de la interv. (randomización: éxito?)	−0.171	−0.10	−0.315	−0.241
$T_{t2} - K_{t1}$		0.7	0.317	0.228	0.286
$T_{t1} - K_{t2}$		−0.13	−0.083	−0.518	−0.379
$T_{t2} - K_{t2}$	diferencias post interv. (efecto de tratamiento)	0.813	0.310	−0.052	0.081
Tt2 – Tt1	con intervención	1.11	0.49	0.59	0.595

```
# Cohens delta R-Code
# group1 = treatment
# group2 = control
# d ~ treatment - control = diff from perspective of treatment
comps.trti <- combn(length(levels(gd)),2) # each 2er comparisons
comps.trti <- t(comps.trti)
cd.treattime <- list()
compnameN <- names(table(gd))
comparisons <- vector()
for(i in 1:nrow(comps.trti))
{
compname <- paste(compnameN[comps.trti[i,2]],
"-minus-",compnameN[comps.trti[i,1]],sep="")
compname
v1.all <- daten.analy[which(gd == compnameN[comps.trti[i,1]]),]
v2.all <- daten.analy[which(gd == compnameN[comps.trti[i,2]]),]
cnamen <- colnames(v1.all)
stopifnot(cnamen == colnames(v2.all))
tmp <- unlist(lapply(seq_along(v1.all),function(x)
{
# perspective > > > Cohens d: v2 - v1 = treat - control
cohensd(v1.all[,x], v2.all[,x])[2]
}
))
cd.treattime[[i]] <- tmp
comparisons[i] <- compname
}
cd.treattime <- (do.call("rbind",cd.treattime))
colnames(cd.treattime) <- c(cnams)
rownames(cd.treattime) <- comparisons
cd.treattime
```

y ahora la salida

```
# C = control group
# T = treatment group
                          SDNNms  RMSSDms  sd2/sd1 log(sd2/sd1)
C.t2.post-minus-C.t1.pre  -0.0464 -0.00806  0.2403  0.1572
T.t1.pre-minus-C.t1.pre   -0.1707 -0.09900 -0.3150 -0.2414
T.t2.post-minus-C.t1.pre   0.6997  0.31665  0.2281  0.2857
T.t1.pre-minus-C.t2.post  -0.1296 -0.08341 -0.5183 -0.3786
T.t2.post-minus-C.t2.post  0.8130  0.31012 -0.0518  0.0806
T.t2.post-minus-T.t1.pre   1.1063  0.48978  0.5899  0.5948
```

La *edad* se incluyó como *variable exploratoria* tras un debate. En esta fase, no había razones teóricas de peso para hacerlo. Esta elección representó el punto de vista de la *no comprobación* de hipótesis. La edad como variable debía utilizarse con cautela si se quería obtener información adicional sobre la distribución por sexos y, por tanto, sobre la propia muestra (véase la Tabla 5.10).

```
> # age x sex structure distribution
> table(sex,age)
age
sex 20 21 23 27 28 29 33 35 36 37 46 50 52 53 57 58
m    0  0  0  0  2  0  2  2  2  4  2  2  2  2  2  2
w    2  2  2  4  0  2  0  0  0  0  0  0  0  0  0  0
```

En pocas palabras, no había ninguna mujer mayor de 29 años en esta muestra y ningún hombre menores de 28 años. Por tanto, había un sistema en la distribución por edades de modo que faltaban mujeres mayores de 29 años y hombres menores de 28. Además, prácticamente no había coincidencia entre los sexos en cuanto a la edad. Un vistazo a la tabla bastaba.

Tabla 5.10: *Estudio Wipfler (2017, distribución por sexo y edad)*

		Edad															
		20	21	23	27	28	29	33	35	36	37	46	50	52	53	57	58
Sexo	m	0	0	0	0	2	0	2	2	2	4	2	2	2	2	2	2
	w	2	2	2	4	0	2	0	0	0	0	0	0	0	0	0	0

No fue necesario realizar una prueba χ^2 para determinar la falta de distribución equitativa y predecir las consecuencias de esta constelación en los resultados. La correlación de la edad y el género fue de r = -0.84, dependiendo el signo únicamente de la polaridad de la variable de género. No existía una distribución realista para la edad y el sexo en esta muestra, salvo quizá para los empleados de un concesionario de automóviles (véase más adelante). Cualitativamente podemos extraer de ello hipótesis prometedoras sobre la todavía válida y ciertamente cuestionable desde el punto de vista de la igualdad. Sin embargo, esta pregunta, que se desvió hacia el debate sobre el género, no tenía mayor propósito para el estudio. Pero ya sabemos que faltan datos importantes y las conclusiones sólo serán posibles hasta cierto punto.

Pasemos a los resultados: Para las variables SDNNms y RMSSDms los resultados se ajustan a las hipótesis formuladas anteriormente y no profundizaremos en ellos. Además, se observó un fuerte efecto de la edad, que – al ser de carácter exploratorio – tendría que ser investigado más a fondo en el futuro. Podría ser el resultado de una variación relacionada con la muestra, como indica el tabla de distribución edad-sexo (véase la Tabla 5.10). El análisis de residuos (Loy, 2021) con el paquete `HLMdiag` de R no reveló incoherencias significativas y las estimaciones de los parámetros se validaron mediante simulación (bootstrap paramétrico) utilizando R.

Si los modelos se utilizaran para la predicción y los β-pesos de los predictores se interpretaran de forma más estrechamente, podría haber tenido sentido eliminar los efectos principales del *tratamiento* y del *tiempo* del modelo. En consonancia con las expectativas, no hicieron una contribución al modelo, ni en términos de dirección ni de magnitud (véanse los errores de tipo S y de tipo M en Gelman & Hill, 2007; Gelman & Hill, 2007; Gelman & Carlin, 2014, véase el capítulo 4.3.3.2). El término de *interacción Tiempo x Trata-*

miento era el aspecto del modelo que realmente interesaba en el presente caso. Este ajuste del modelo no era relevante en el presente caso. Se trataba simplemente de establecer que el tratamiento se ajustaba a la hipótesis y que la aleatorización funcionaba (es decir, que no había diferencias sustanciales en el momento t1/ Pre). Una interpretación del contenido de los predictores o una predicción basada en los β-pesos en relación con los cambios en las escalas subyacentes de las variables dependientes no era una cuestión de este estudio y, por lo tanto, no se llevó a cabo.

Debido a las *altas intercorrelaciones* entre *SDNNms* y *RMSSDms* (véanse la Fig. 5.34 o la Tabla 5.8), no es sorprendente que los modelos lineales jerárquicos calculados para estas variables condujeran a resultados muy comparables. Un vistazo a la tabla de intercorrelaciones de todas las variables potenciales (no impresa) – después de todo, el instrumento de medición para la VFC elaboró k = 23 variables – muestra valores desde r = 0.002 hasta r = 0.99. Si aquí se seleccionaron variables sin ninguna base teórica (véase la Fig. 5.35), esto podría conducir a una abundancia incontrolada de resultados estadísticos *altamente significativos* sin una gran ganancia de conocimiento, aunque, sobre todo, se investigaría más o menos lo mismo y, por tanto, no se obtendría ningún conocimiento. En principio, sería posible realizar un análisis de componentes principales sobre esta avalancha de variables y luego continuar con los valores factoriales de los – presumiblemente – pocos factores resultantes. Sin embargo, cabe preguntarse qué se investigaría exactamente. Aquí el fabricante de los instrumentos debería intervenir, ya que no se puede suponer que los profesionales de la quiropráctica disfrutaran al mismo tiempo de una sólida formación científica en métodos de investigación. Es engañoso que todas las variables se denominen de forma diferente y se asocien a conceptos muy diversos. Pero si una variable se correlaciona con otra a casi r = 0.99, ya no se trata de dos variables o conceptos estructuralmente diferentes. La figura 5.35 muestra un correlograma de todas las variables que el instrumento de medida mostró y del que se desprende fácilmente el problema, independientemente del hecho de que las variables sean en parte bipartitas o muestren un claro sesgo a la derecha, dada la muestra inestable, muy unilateral y pequeña,. En vista de 23 variables, que producen (23^2-23)/2 = 253 correlaciones, uno se pregunta seriamente por qué la estadística clásica calcula los valores *p* y qué se supone que nos dicen. A un nivel del 5% esperamos 253*.05 = 13 correlaciones "significativas", pero dependiendo de la variación de la muestra, sin embargo, no sabemos qué correlaciones son las *reales* y cuáles no y, sobre todo, qué ocurre en la *zona gris*. No sólo por esta razón tiene sentido limitarse a unos pocos valores teóricamente interesantes, olvidar el resto y no ceñirnos a la autenticidad de correlaciones con valores *p*, sino utilizarlas como una fuente de tantas informaciones.

```
# all data
pal <- terrain.colors(4)
# diaglabel <- as.character(abbreviate(names(rawd),labbrev))
corrgram(daten.cor, lower.panel="panel.ellipse", upper.panel="panel.pie",
          diag.panel="panel.density", col.regions=colorRampPalette(pal),
          main="Corrgram", order=FALSE)
```

El análisis de la relación SD2/SD1 fue diferente del de SDNNms y RMSSDms. En primer lugar, se observó que, como variable de relación, no se distribuía realmente de forma normal, sino que estaba muy sesgada (véase la Fig. 5.36). Como puede verse, una transformación logarítmica aportó una mejora – más que la simple extracción de raíces. Las transformaciones de datos no lineales, como la logaritmización, se tratan de forma diferente en la bibliografía (por ejemplo, Gelman, 2015c, Matechou, 2013-12-04, pero véase Feng, 2014), desde "sí, simplemente" hasta "no, simplemente no, ¡por el amor de Dios!". Lamentablemente, se carece de una justificación autorizada e independiente del contexto para cualquiera de estas posturas. Si acudimos al original de Tukey (1977), tanto la transformación `log()` como la transformación de datos con raíces cuadradas desempeñan un papel relevante para él (véase el estudio de caso anterior, cap. 5.5.1). El propio Tukey (1977, p.153, negrita en el original) hace observaciones al respecto,

„In this example, we have again seen the same main points as in the previous one:
- **changing scales to make dependences roughly linear usually helps.**
- **attening by substraction makes it much easier to see what is going on**

at the more subtle levels.
[...]
- **the uselfulness of changing scales to reduce confusion caused by crowding.“**

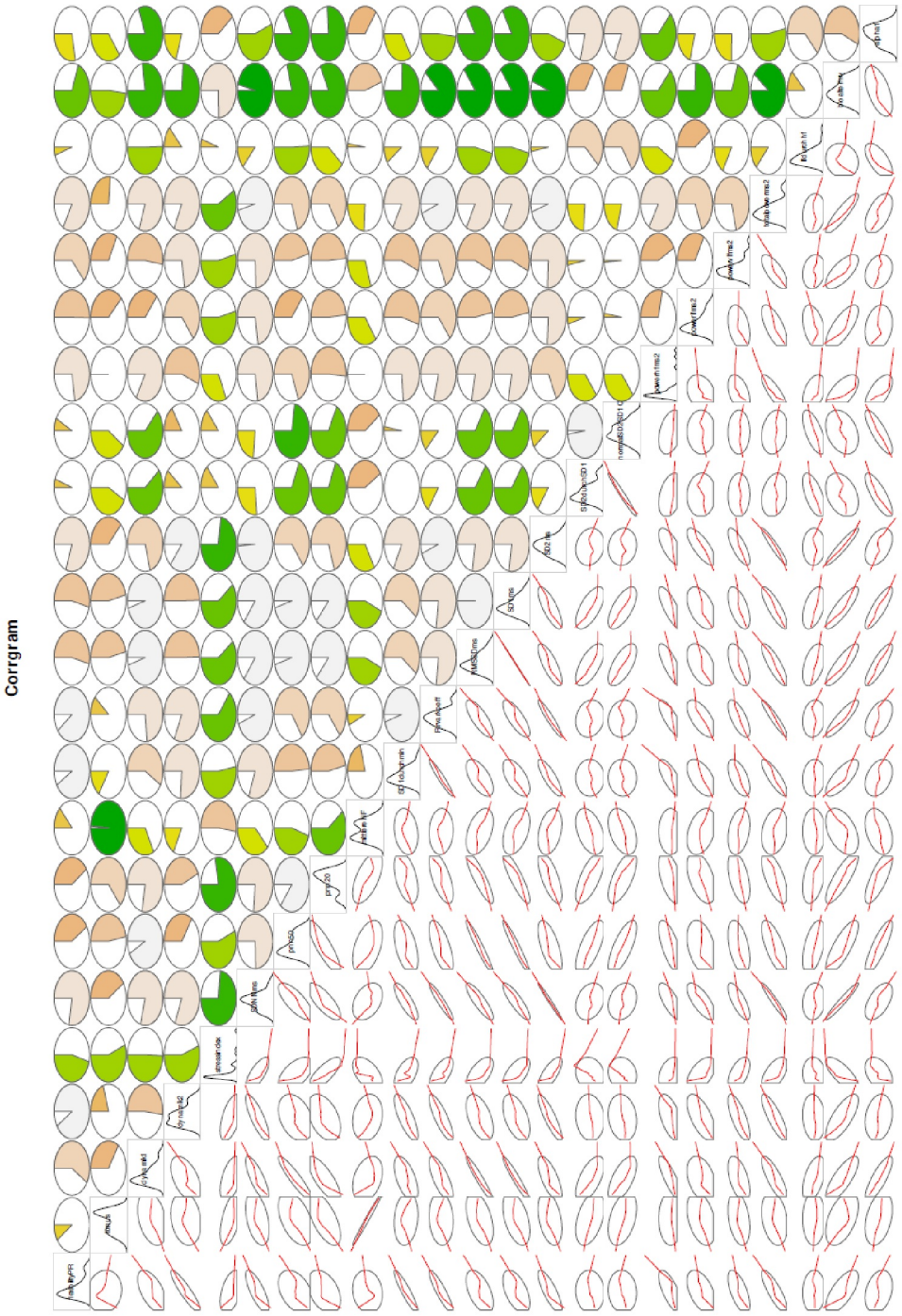


Figura 5.35. *Estudio Wipfler (2017, correlograma de todas las variables)*

Por tanto, parecía perfectamente legítimo transformar los datos con reservas y ser conscientes de que las transformaciones no lineales modifican las estructuras de los datos. La "distorsión" nos parece una descripción inadecuada de este proceso, ya que el cambio es deseado. Y la pregunta sería "¿distorsionado hacia qué?" de todos modos. Porque se podría argumentar que los datos están realmente en una escala logarítmica y no como se recogen empíricamente. Entonces, la transformación de los datos sólo los cambiaría de forma que volvieran a acercarse a su nivel original. Esto significaría que los datos se recogieron de forma distorsionada y se igualaron mediante la transformación. Entonces, ¿cuál es la verdadera distribución de los datos? Los instrumentos condicionan la naturaleza de los datos, por lo que siempre podría ser diferente.

Además, suponiendo que las variables individuales SD1 y SD2 se distribuyan cada una normalmente, su cociente sigue en principio una distribución de Cauchy, que no conoce ni la media ni la desviación estándar. Sin embargo, dado que aquí se trata de un intervalo de valores reales muy estrecho y limitado por naturaleza, parecía justificado trabajar con los procedimientos descritos tras la transformación de los datos y ser conscientes de estas limitaciones. Sin embargo, esto requiere que el análisis se lleve a cabo tanto de esta forma como de otra. En el caso concreto, esto significa *con y sin* transformación `log()`. De las diferencias aprendemos algo sobre los datos, lo que sin duda es exploratorio en el sentido de AED. Sin esta exploración, se aplicaría de forma generalizada una lógica de fiabilidad de los signos sin tener en cuenta la información contextual y la ganancia de conocimiento sería significativamente menor.

```
# histograms for sd2/sd1
# comparison no transfo, log(), sqrt()
par(mar=c(5,5,4,2), oma=c(2,1,1,1), cex.axis=0.8, mfrow=c(2,2))
hist(daten.analy[,"sd2/sd1"], prob=TRUE, main="", xlab="sd2/sd1",
           col="skyblue", border="white", pre.plot=grid())
lines(density(daten.analy[,"sd2/sd1"]), col="violetred2", lwd=2, lty=2)
sd2.sd1.log <- log(daten.analy[,"sd2/sd1"])
hist(sd2.sd1.log, prob=TRUE, main="", xlab="log(sd2/sd1)",
           col="skyblue", border="white", pre.plot=grid())
lines(density(sd2.sd1.log), col="violetred2", lwd=2, lty=2)
sd2.sd1.sqrt <- sqrt(daten.analy[,"sd2/sd1"])
hist(sd2.sd1.sqrt, prob=TRUE, main="", xlab="sqrt(sd2/sd1)",
           col="skyblue", border="white", pre.plot=grid())
lines(density(sd2.sd1.sqrt), col="violetred2", lwd=2, lty=2)
mtext("Study Wipfler / chiropractic (SD2/SD1)", 3, line=-1, cex=1.5, outer=TRUE)
mtext("Histogram and density", 3, line=-2.6, cex=1.1, outer=TRUE)
```

El modelo lineal de *SDNNms y RMSSDms* (véase más arriba) demostró ser superior a un modelo reducido (anotación R) – por ejemplo, en términos de errores estándar más pequeños:

```
log(SD2/SD1) ~ Zeitpunkt * Alter + (1|Subjekt)
```

En este modelo, el *tratamiento* uV ya no desempeñaba ningún papel. Aparentemente, el *tratamiento* no influía. Esto, a su vez, hizo que la variable resultara interesante para futuras investigaciones exploratorias. ¿Por qué el tratamiento no tuvo el efecto esperado? Por lo tanto, se abandonó el campo de los enfoques hipotético-deductivos puros.

El *efecto de interacción del tiempo y la edad* significaba inicialmente que, en función de la edad, el *curso* (t1/ Pre a t2/ Post) de la variable *log(SD2/SD1)* era diferente, pero que la edad no ejercía una influencia única independiente. Según la tabla 5.10 anterior, estaba claro que la composición desfavorable de la muestra producía un entrelazamiento inseparable de la edad y el sexo. Así, prácticamente todas las mujeres eran más jóvenes que los hombres y casi no había superposición de sexo y edad. Esta distribución reflejaba ante todo la política de personal del sector automovilístico. Por lo tanto, una tesis no especialmente atrevida aparte del diseño fue – las mujeres (jóvenes) trabajan en recepción y back-office, los hombres (mayores) en ventas y consultoría o dirección. Hay una rigurosa separación de sexos. En sentido estricto, ambas variables podrían o deberían haberse fusionado en una nueva variable edad-género, lo que habría dado lugar a mayores complicaciones y no habría facilitado realmente una interpretación.

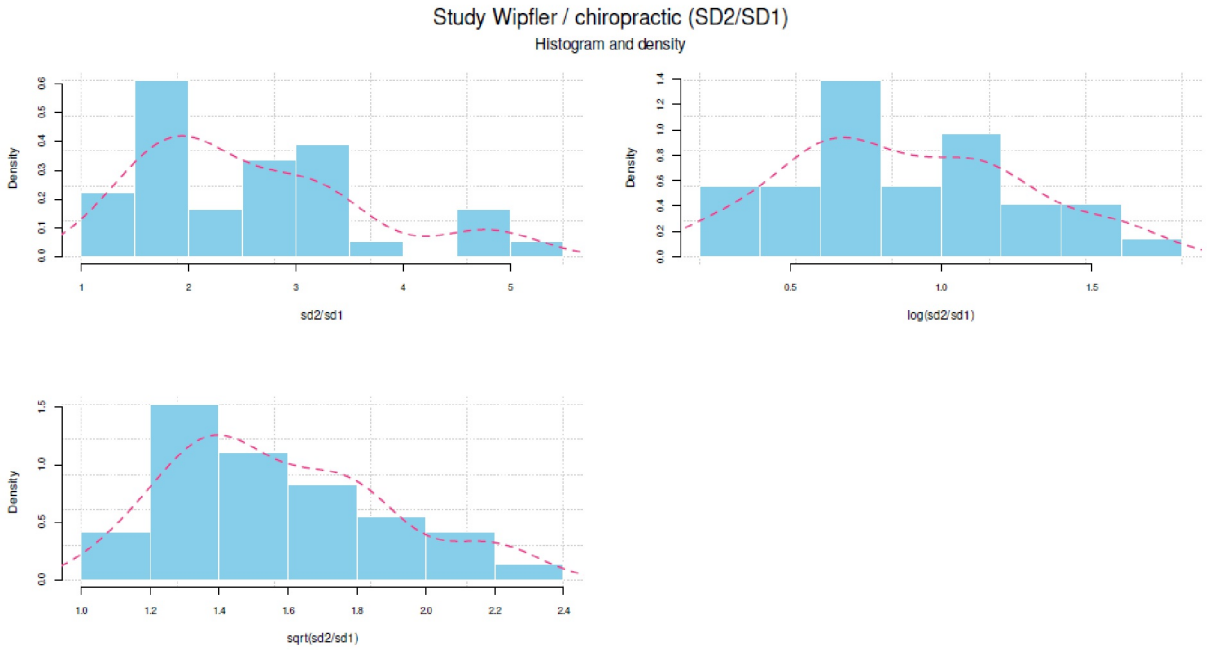


Figura 5.36. *Estudio de Wipfler (2017, SD2/SD1, histogramas)*

Además, hubo un *efecto principal* para el momento, lo que significa que, en contra de lo esperado, ambos grupos cambiaron significativamente con el tiempo. Sin embargo, no estaba claro cómo y por qué se producían estos cambios. Esto debía investigarse con ayuda de AED. Para ello, se generó un gráfico de interacción en diferentes variaciones (véase la Fig. 5.37), que representaba la variable a lo largo del tiempo y para los grupos por separado, *con y sin* transformación `log()`. Para la salida del gráfico de interacción escribimos un `ia.plot()`, de modo que el diseño ya está adaptado a nuestras necesidades. Además los valores característicos se abrevian de antemano para que el gráfico sea más legible.

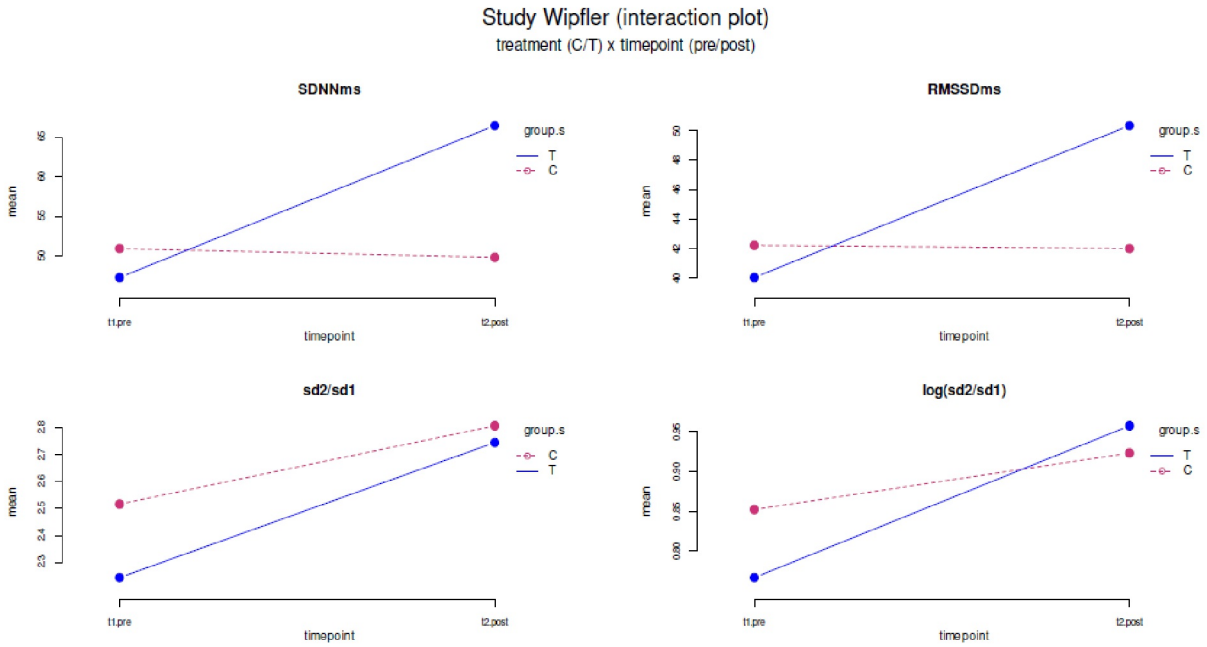


Figura 5.37. *Estudio de Wipfler (2017, gráficos de interacción aVs, punto temporal * tratamiento).*

```
# vars and specs
colo <- c("violetred3","blue","yellowgreen","orange")
group.s <- factor(group, labels=c("C","T"))
time.sex <- paste(timepoint, sex, sep=" | ")
vars <- c("SDNNms","RMSSDms","sd2/sd1","log(sd2/sd1)")
facs <- data.frame(timepoint,sex,group,group.s,time.sex)
TITLE <- "Study Wipfler (interaction plot)"

# by treatment x timepoint
SUB <- "treatment (C/T) x timepoint (pre/post)"
ia.plot(dframe=daten.analy, facs=facs, f1nam="timepoint",
           f2nam="group.s", vars=vars, TITLE=TITLE, SUB=SUB,
           colo=colo, trace.label="group.s")
```

En primer lugar, se observó que el grupo de control tenía *un nivel más alto* que el grupo de tratamiento inicialmente en el momento t1/pre. Como sabíamos por las otras variables que esto no siempre es necesariamente así y la asignación fue aleatoria, esto podría deberse a una variación natural, lo que los estadísticos clásicos llaman de forma inespecífica un *artefacto de muestreo aleatorio*. No queremos sobreinterpretar nada, pero no nos gusta mucho el concepto de aleatoriedad, ya que simplemente oculta el hecho de que las relaciones causa-efecto siguen funcionando, sólo que más allá de nuestro conocimiento. Además, hubo un *aumento pronunciado* con el tratamiento y un *aumento menos pronunciado* con el control. Visto de forma aislada, podría concluirse que el tratamiento había funcionado aquí con la misma fiabilidad, pero no sólo, sino en el contexto de una amplia gama de variables e influencias. Llegados a este punto, tendríamos que preguntarnos qué influencias aparte del tratamiento podrían cambiar sustancialmente el valor log(SD2/SD1). Si se consideraba la variable SD2/SD1 sin la transformación log() (abajo a la izquierda en la Fig. 5.37), se observaba que la descripción era similar a la de los valores `log()` similar a los datos transformados con `log()` (parte inferior derecha de la Fig. 5.37). Lo que era diferente, sin embargo, era que los dos cursos no se cruzaban y el tratamiento no tenía valores más altos que el control en el tiempo t2/post. Asumiendo que la transformación `log()` fue la elección correcta aquí, el gráfico nos mostró que el efecto de la intervención quiropráctica era más evidente con la transformación `log()` de acuerdo con la hipótesis. Comparemos esta imagen con los gráficos de interacción de las variables SDNNms y RMSSDms, los patrones eran básicamente los mismos, sin entrar en las cifras exactas.

Estaba claro que en la presente muestra muchas variables interactuaban y no estaba claro si todas estaban relacionadas con la intervención quiropráctica. Por esta razón, se examinó experimentalmente un modelo complejo de la forma

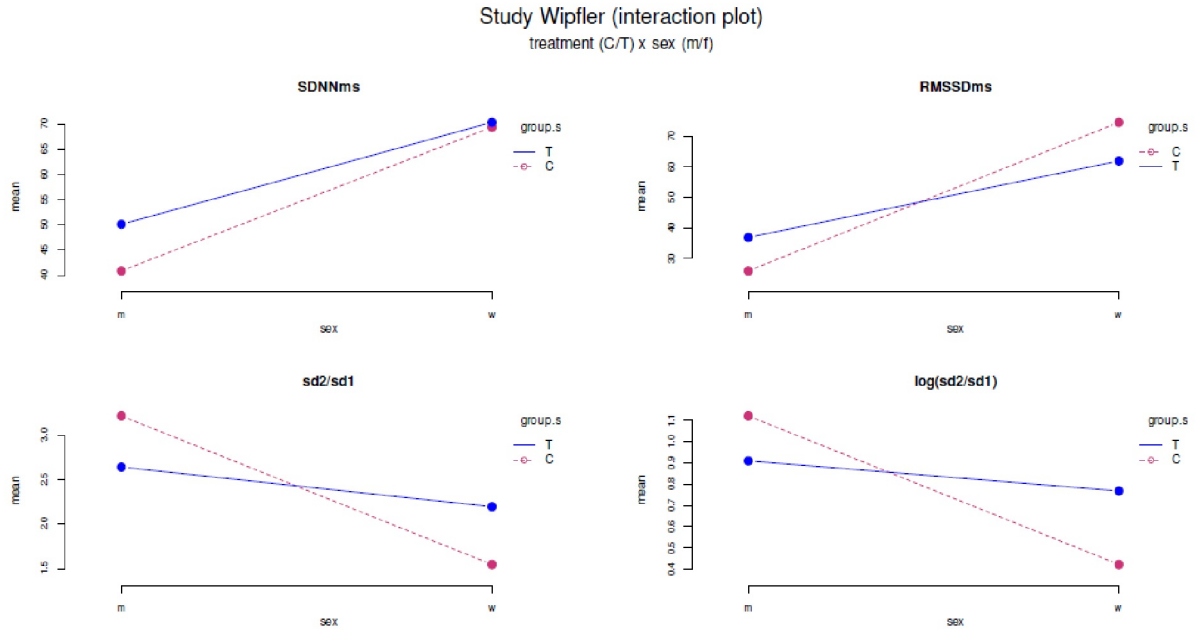
```
log(SD2/SD1) ~ Tratamiento * Punto temporal * Edad * Sexo + (1|Subjecto)
```

Esto demostró ser superior a la primera solución (véase más arriba) en la prueba de Likelihood Ratio (aquí: comparación de ambos modelos) mediante `anova()` en R y según los índices de calidad comunes de los modelos lineales jerárquicos, y contenía cierto poder explicativo ampliado con respecto al contexto que nos ocupa. Por otra parte, dado n = 9 personas, k = 2 grupos y t = 2 puntos temporales (= 36 puntos de datos), nos parecería atrevido interpretar dicho modelo con demasiado detalle: ¿Quién seguiría entendiendo correctamente una interacción de cuatro vías dados los pocos puntos de datos disponibles (personas, tiempo)? En realidad, lo único que significaba era que la influencia de la edad y el sexo influía mucho en el curso de los grupos, más de lo que podría hacerlo el tratamiento por sí solo. Un análisis de comparación individual post-hoc más preciso utilizando `lsmeans()` del paquete `LSMEANS` de R no mostró

- ningún efecto principal sustancial
- Tratamiento * Momento (t1/ Pre tras t2/Post cada uno para control y tratamiento)
- Punto temporal * Sexo (curso hombres t1/pre tras t2/post, pero no para las mujeres).

Explorando, merecía la pena volver a los gráficos de interacción con esta información:

1. Tratamiento x Tiempo – véase Figura 5.37
2. Sexo x Tratamiento – véase Figura 5.38
3. Sexo x Tiempo – véase también el entrelazamiento desfavorable de la edad y el género, véase la Figura 5.39.
4. (Punto temporal x Sexo) Tratamiento – véase Figura 5.40



Figur 5.38. *Estudio de Wipfler (2017, gráficos de interacción aVs, tratamiento * sexo).*

Los diferentes gráficos de interacción de grupo, sexo y punto temporal en las distintas combinaciones muestran los cambios y diferencias relevantes entre los uV, así como cualquier interacción. Los métodos numéricos deberían entonces proporcionar únicamente los valores numéricos y coeficientes asociados. Por lo tanto, preferimos fijarnos en los gráficos en lugar de en los coeficientes, ya que suponemos que los efectos relevantes deben ser visibles en un gráfico y que el análisis numérico posterior sólo confirma este esta visión y la hace numéricamente más precisa y muestra los cambios relevantes y las diferencias entre los uV, así como cualquier interacción. Se puede comparar a las tablas del conjunto de datos del Titanic (véase más arriba), también es posible con los gráficos de interacción se pueden relacionar combinaciones de variables y sus características de forma combinatoria.

```
# by sex x treatment
SUB <- "treatment (C/T) x sex (m/f)"
ia.plot(dframe=daten.analy, facs=facs, f1nam="sex", f2nam="group.s",
           vars=vars, TITLE=TITLE, SUB=SUB, colo=colo, trace.label="group.s")

# by sex x timepoint
SUB <- "sex (m/f) x timepoint (pre/post)"
ia.plot(dframe=daten.analy,facs=facs, f1nam="timepoint", f2nam="sex",
        vars=vars, TITLE=TITLE, SUB=SUB, colo=colo, trace.label="sex")

# by (timepoint x sex) x treatment
SUB <- "[timepoint (pre/post) x sex (m/f)] x treatment (C/T)"
ia.plot(dframe=daten.analy, facs=facs, f1nam="group", f2nam="time.sex",
        vars=vars, TITLE=TITLE, SUB=SUB, colo=colo, trace.label="time.sex")
```

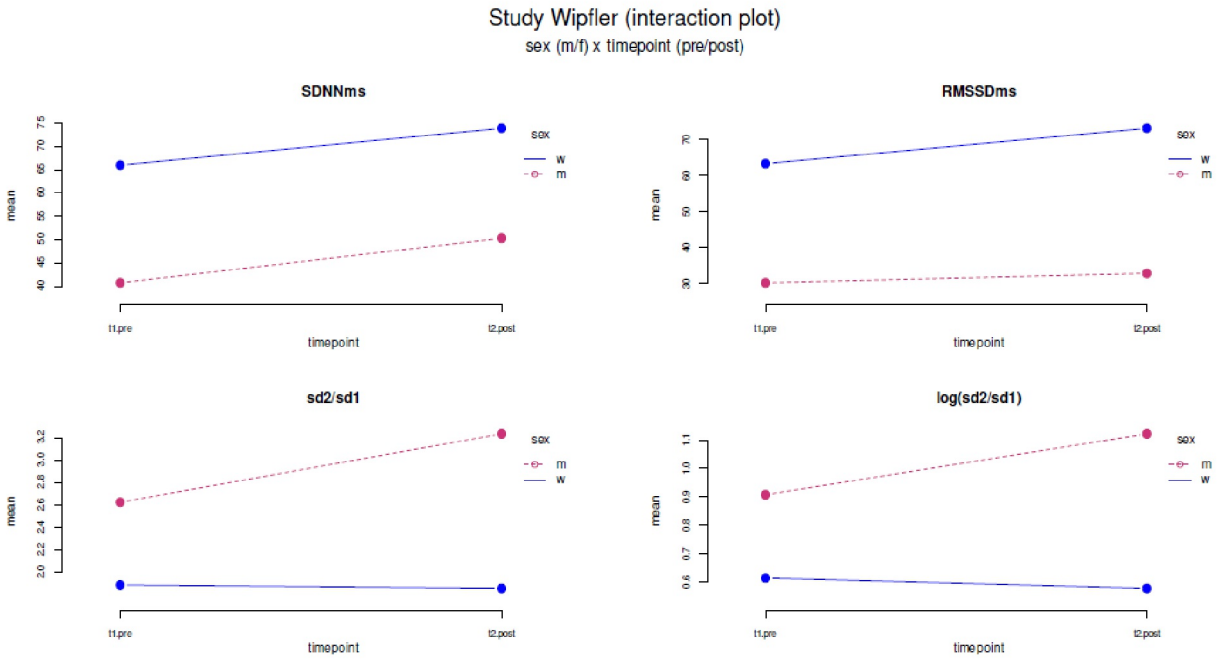


Figura 5.39. *Estudio de Wipfler (2017, gráficos de interacción aVs, punto temporal * sexo).*

Como puede observarse, los grupos se comportaron de forma diferente en función del momento temporal y el grupo de control, en particular, mostró un cambio a lo largo del tiempo, que ya se ha mencionado anteriormente. Además, los cambios temporales funcionaron para los hombres, pero no para las mujeres. Sin embargo, esto sólo fue cierto en parte. Como muestra la Figura 5.40, aunque los hombres se beneficiaron más del tratamiento en log(SD2/SD1), esto podría deberse a que los hombres eran significativamente mayores en la muestra. Y no fue así en el caso de las otras variables SDNNms o RMSSDms, en las que las mujeres parecieron beneficiarse más, como puede verse en la Figura 5.40. En este punto, Wipfler consultó la bibliografía sobre la VFC y la cuestión de la edad y el sexo. Tras investigar encontró indicios de que existen efectos tanto de género como de edad sobre la VFC que deberían investigarse y debatirse con más detalle. Así pues, no se encontró ninguna variación puramente relacionada con la muestra que ocultara el efecto real, sino más bien influencias adicionales que habría que tener en cuenta en el futuro de forma mucho más exhaustiva y precisa. Posteriormente, si fuera posible, en un estudio de replicación deberían incluirse *random intercept y slope* para el sexo y la edad, aparte de múltiples puntos de medición y múltiples intervenciones quiroprácticas. Del mismo modo, sería necesaria una selección de variables teóricamente más sólida y una discusión seria con el fabricante de los instrumentos de medición, que obviamente producen sobre todo redundancias. El número de variables de medición relevantes tendría que reducirse significativamente en el futuro a unas pocas variables robustas y teóricamente sólidas. Un futuro estudio podría tener un diseño similar al siguiente. Una nueva encuesta en un contexto comparable (aquí: concesionario de coches) está descartada. La futura selección de la muestra tendría que planificarse con mucho cuidado.

Consideramos que todos estos son los resultados significativos de este estudio: una identificación relativamente precisa de los factores que influyen y suficientes hipótesis sobre cómo podrían interactuar para orientarlo de forma muy específica la próxima vez. No obstante, dada la estrecha asociación de la edad y el sexo en esta muestra, los resultados no deben interpretarse de forma exagerada. Es importante señalar que, en el modelo más complejo, el efecto de interacción de interés, tratamiento x punto temporal volvió a aparecer, que apoyaba el tratamiento – pero aquí también era válido para el grupo de control – y no aparecieron otros efectos principales. No discutimos los análisis residuales y la identificación de valores atípicos, que también se llevaron a cabo y no produjeron ningún hallazgo significativo que hubiera cuestionado los resultados descritos. Así pues, finalizamos el análisis y podemos pasar a la discusión de la contribución del AED.

5.5.6.1 ¿Qué modelo es mejor?

En relación con la variable *log(SD2/SD1)*, podemos afirmar claramente que *ninguno es el mejor modelo*. Sin embargo, nos parece que el más plausible es el más complejo, ya que parece congruente con los resultados de la bibliografía. Sin embargo, dada la composición y el tamaño de la muestra, seguimos sin querer interpretarlo. El AED y sobre todo los gráficos de interacción nos han permitido diferenciar las distintas variables influyentes para esta variable. Si tenemos en cuenta el diseño, la distribución por sexo/edad y los modelos HLM/ MLM, debería quedar claro que un procedimiento puramente hipotético-deductivo basado en una lógica de significación no habría podido aportar estas conclusiones. Un enfoque puramente probatorio se habría detenido en el valor *p* y el tratamiento sólo se habría clasificado como eficaz o no eficaz y, dada la incertidumbre en la determinación de los valores *p* en los HLM, éstos no son muy relevantes. De hecho, su cálculo está lejos de ser inequívoco, como Bates (2006), el autor de `lmer()`, argumenta en detalle (véase también Bolker et al., 2020). Así pues, el cálculo de los grados de libertad no se ha aclarado y, por lo tanto, las conclusiones basadas en los valores *p* se consideran erróneas en principio, ya que son poco claras e imprecisas. Las desviaciones a través de la AED han demostrado que la futura consideración por separado de la edad en la VFC podría ser un complemento útil. Para cada una de ellas, podrían modelizarse diferentes niveles y cursos en el marco de un HLM/MLM. Esto no fue posible en el presente estudio debido a la escasa cantidad de datos. Asimismo, el AED proporcionó argumentos bastante convincentes de que el tratamiento funcionaba de acuerdo con la hipótesis. Así pues, el ajuste de modelos y el AED van de la mano.

5.5.6.2 ¿Qué aprendemos de todo esto?

En primer lugar, aprendemos la importancia de la selección de la muestra y una mirada al diseño. Si sólo queremos estudiar los concesionarios de coches, podemos tener una muestra muy representativa. Si, por el contrario, queremos hacer afirmaciones un poco más amplias, es obvio que los hombres pueden ser menores de 28 años y las mujeres mayores de 29, y en función de eso probablemente difieran en otras características que no se han estudiado en absoluto. Y los datos que se examinaron específicamente no dicen nada al respecto. Los factores limitantes son la composición de la muestra (edad, sexo), el pequeño tamaño de la muestra y la falta de mediciones múltiples.

Si creyéramos únicamente en coeficientes como los valores *p* y no examináramos los datos cualitativamente y gráficamente en función de las condiciones contextuales, pasaríamos por alto el hecho de que el tratamiento probablemente fue eficaz para todas las variables. Pero estaba enmascarado por otros factores, dependiendo de la variable que parece haber tenido un efecto sobre la variable *log(SD2/SD1)* pero no en la misma medida sobre las otras dos (*SDNNms, RMSSDms*). Estos factores desconocidos deberán encontrarse en el futuro. Del mismo modo, los gráficos indican que la transformación `log()` mostró efectos más claros en este caso, incluso si – en términos de valores *p* – no había realmente una diferencia entre *SD2/SD1* y *log(SD2/SD1)*. Esto se investigó adicionalmente, sin informarlo en detalle aquí. No obstante, se utilizaron los datos transformados `log()`. En primer lugar, porque estaba bien justificado desde el punto de vista del contenido y en segundo lugar porque los cambios esperados a lo largo del tratamiento y, por tanto, toda la situación de los datos parecían más coherentes. En el caso de variables potencialmente significativas, es importante no lanzarse a ciegas sobre ellas, sino plantearse "si tienen un efecto, ¿por qué lo tienen?

Surgieron preguntas abiertas para los efectos exactos de la edad y el sexo, así como con respecto al curso a largo plazo no investigado de la intervención quiropráctica y sus efectos sobre la VFC. Aquí sería necesario repetir el experimento con el marco ampliado de variables relevantes (edad, sexo), con varios puntos temporales de medición e intervención quiropráctica y una muestra representativa. Por lo tanto, somos extremadamente cautelosos a la hora de hablar de efectos de la edad y el sexo, aunque existan pruebas de ello.

5.5.6.3 ¿Dónde está aquí la integración de métodos?

La integración cuantitativa-cualitativa de métodos en el sentido de métodos mixtos no existía en el presente estudio en este punto. El objetivo de la descripción de este estudio era dar a conocer las múltiples formas en que el AED puede producir resultados. Sin embargo, la discusión y el análisis necesarios siguieron aspectos muy marcadamente cualitativos, de modo que el aspecto de los métodos mixtos se añadió por la puerta de atrás. Consideramos que este intento tuvo éxito porque se pudo reconstruir de forma comprensible la interacción de las variables implicadas. Y esto, a su vez, fue el resultado de un análisis cualitativo centrado en la información cuantitativa disponible sin cálculos adicionales. Además, surgieron directrices claras para la repetición con variables influyentes.

Sin embargo, donde un estudio cualitativo complementario se hace imperativo en el contexto de la quiropráctica es en los efectos concretos sobre las prácticas vitales de las personas. Entre ellas se incluyen los comportamientos de salud, el ejercicio y el deporte, el bienestar general y la satisfacción, la nutrición, los problemas de espalda, el dolor y el uso de analgésicos y – de forma más general – el impacto en los costes sanitarios individuales. Aquí valdría la pena explorar las biografías típicas de los pacientes para identificar los puntos de partida de una autonomía relativamente mayor. Especialmente en el contexto de las deficiencias físicas graves, la medicina convencional tiende a cronificar a los pacientes, lo que va acompañado de una drástica disminución de la autonomía experimentada y vivida, posiblemente con una actitud de sufrimiento pasivo, lo que puede conducir a un círculo vicioso. Además de los datos sobre, por ejemplo, el consumo de analgésicos, la actividad física, tratamientos y terapias necesarios, etc., el análisis cualitativo puede ayudar a comprender el punto de vista subjetivo, reconstruir las acciones y decisiones en su respectivo contexto y elaborar cursos típicos. A partir de ahí, se pueden derivar recomendaciones concretas para los pacientes, formación y perfeccionamiento para los profesionales, etc., y evaluar su eficacia. A largo plazo, el repertorio de médicos y remisores y políticamente la asunción de costes por parte de las compañías de seguros sanitarios, etc. pueden modificarse. Para ello se requiere una teoría bien fundamentada y datos empíricos que apoyen la teoría.

5.5.7 Potenciales de recuperación en la terapia de la adicción

Entre 1992 y 1998, U.M. Studer (1998) llevó a cabo una evaluación a largo plazo del centro suizo de terapia de la adicción *start again* de Zúrich, financiado por la Oficina Federal de Justicia de Suiza (BAJ). El diseño consistió en una combinación de estadística de Bayes para muestras pequeñas (Bretthorst, 1993) y análisis de secuencias (véase el capítulo 11.5) para la reconstrucción de casos ejemplares de adictos y sus biografías. Sobre esta base de datos se difundieron de nuevo los métodos de trabajo y los efectos del enfoque terapéutico de la *sistémica profunda* practicada en el inicio. En Gürtler, Studer y Scholz (2012), se examinan con más detalle cinco historias de casos catamnésicos después de que los respectivos clientes abandonaran *start again*. Se realizaron tres entrevistas muy abiertas cada una en relación con los acontecimientos posteriores al abandono de la terapia. Dos casos de esta muestra ya desempeñaron un papel importante en el estudio de Studer (1998), por lo que se disponía de dos historias a largo plazo muy bien documentadas para el estudio. Las catamnesis fueron analizadas secuencialmente para evaluar el curso después de la terapia y así estimar el potencial de integración con respecto a la adicción relativa aún vivida o no. Tres entrevistas cada una con los expedientes completos de los casos del periodo de terapia sirvieron como base de datos. Éstos contenían diferentes informaciones. Entre ellos se incluyen los genogramas y en algunos casos las cartas de solicitud de plaza en la terapia, que en su mayoría fueron escritas desde el síndrome de abstinencia en el psiquiátrico. Esto se completó con notas sobre el transcurso de la terapia y con "puntos de vista" (= auto-reflexiones) escritos por los propios clientes sobre su propio desarrollo. El objetivo del análisis era, por un lado, evaluar el potencial de integración potencial de integración, la derivación de retos para las respectivas personas en el momento actual y, por otro, en un nivel abstracto derivar un modelo paso a paso para la reconstrucción de un camino exitoso para salir de la adicción (Gürtler, Studer & Scholz, 2012).

Tabla 5.11: *Estudio de catamnesis, Gürtler, Studer y Scholz (2012, potencial de recuperación)*

	N. Lang	B. Kaiser	J. Nagel	S. Gabler	N. Widmer
∑ Vulnerabilidades	7	13	13	16	13
∑ Recursos	6	5	5	3	7
Potencial de recuperación = Recursos/Vulnerabilidades	0.86	0.38	0.38	0.19	0.54

A continuación, sólo compararemos el nivel individual de integración entre los casos. Para ello, se compararon las vulnerabilidades (factores de riesgo) con los recursos individuales de los cinco clientes. Las tablas correspondientes contienen los factores de riesgo y los recursos antes de la terapia (Gürtler, Studer & Scholz, 2012, cap. 9.2, tab. 9.1, los nombres son seudónimos) y el curso catamnésico después de la terapia (ibíd., tab. 9.2). En cada caso, se determinó la presencia de los factores de influencia identificables por cliente, se sumaron y, a continuación, se formó la relación entre recursos y vulnerabilidades para evaluar el potencial de recuperación.

Esto presupone varios supuestos, como que, en principio, los recursos y las vulnerabilidades se influyen mutuamente. La suposición va más allá y consiste en que pueden influirse mutuamente en cuanto a su influencia y grado de eficacia, de tal manera que los factores pueden minimizarse o incluso anularse. Otro supuesto es que pueden añadirse nuevas vulnerabilidades y recursos en el transcurso de la vida, ya sea mediante una adquisición activa, acontecimientos vitales críticos u otros factores que se especificarán con más detalle. La tabla 5.11 muestra un extracto de la tabla 9.1 (ibíd.), a saber, una simple evaluación exploratoria del potencial de recuperación antes de entrar en terapia.

El potencial de recuperación (PR) como relación entre los recursos y los factores de riesgo es PR = 1 si los recursos y los factores de riesgo se equilibran entre sí. Si es PR < 1, predominan los complejos problemáticos. Si es PR > 1, predominan los recursos. En consecuencia, se predice que la evolución posterior será favorable o desfavorable. Cuando estos valores y el orden resultante de los antiguos clientes se compararon con los resultados de las reconstrucciones de casos, los resultados fueron bastante similares, pero aun así solo coincidieron en detalle hasta cierto punto (Gürtler, Studer & Scholz, 2012). Una de las principales razones de esta discrepancia era que el desarrollo de los antiguos clientes era diferente y el tiempo transcurrido antes de la terapia no era suficiente para reflejar adecuadamente el estado de cada entrevistado en el momento de la entrevista catamnésica y para describir y explicar su estado diferentes números de años después de la terapia. Así pues, hubo que añadir los *factores durante y después* de la terapia, lo que se no hizo sistemáticamente en el estudio de Gürtler, Studer y Scholz (2012), sino que se documenta en extractos en la tabla 9.2 (ibíd.). Dejamos fuera la duración tras la terapia y la edad como incontrolables. Sin embargo, ambos pueden tener un efecto fundamental. Por ejemplo, puede ocurrir que, tras una determinada historia, algunas cosas sólo sean posibles al cabo de muchos años y que en el momento de la encuesta aún no se hubiera alcanzado ese punto en el tiempo. Aunque entonces se produzca un cambio significativo, será difícil demostrar a qué factores concretos se debe. ¿Es la propia terapia, es la edad, son los acontecimientos de las últimas décadas, pero sólo los posteriores al abandono de la terapia – o se trata de un complejo conglomerado de todas estas influencias? Como señala Studer (1998) en su informe final, el tiempo une. El tiempo bien empleado tiene un positivo.

Los casos pueden ordenarse de forma sencilla sin tener que entrar en estadísticas inferenciales. Para simplificar las cosas, los factores individuales se ponderaron por igual. A posteriori, resultó problemático que no fuera posible determinar un *límite crítico* del potencial de recuperación tal como se acaba de describir con PR = 1 o mayor y menor que Uno se requerían supuestos mucho más precisos de los que se dieron. En sentido estricto, el número de factores examinados y la ponderación de los factores de riesgo y los recursos deberían corresponderse entre sí de tal manera que el citado límite de uno pareciera tener sentido. En un nuevo análisis realizado por los autores en el libro, esto parecía prácticamente imposible. Habría requerido un espacio completo de posibilidades de factores de entrada y sus pesos, así como de relaciones causa-efecto. Esto requiere tanto la visión específica del caso como la visión transversal paralela para integrar dialécticamente ambas con el fin de pesos individuales con la posibilidad de comparación interindividual a una escala razonable y robusta. La individualidad y la comparabilidad de los casos son igualmente

importantes. A este respecto, en un reanálisis de los datos de Gürtler, Studer y Scholz (2012) para este libro, se abandonó por completo. En su lugar, la relativa comparación entre los antiguos clientes es ahora suficiente para formar un ranking – bajo el supuesto de que los factores sumados (tanto los riesgos como los recursos) se escalen por intervalos como factores, lo que permite la formación de un ratio. La comparación relativa es suficiente, ya que una comparación absoluta no tiene ningún criterio razonable. La comparación relativa sólo requiere que cada persona sea responsable de todos los factores considerados eficaces y que las mismas reglas se apliquen a todos ellos.

Surgieron más problemas en el reanálisis, cuando se compararon las categorías de contenido antes de la terapia con las pocas durante la terapia y muchísimas post-hoc a la terapia (= periodo de catamnesis). (véase también la tabla 5.12, pág. 456 sobre la ampliación del potencial de recuperación). Esto se debió a que el historial del caso era muy detallado, las entrevistas sobre la catamnesis también fueron muy detalladas, pero la información sobre el periodo de terapia propiamente dicho fue más bien modesta en comparación con los otros dos periodos. Por eso, tenía sentido tratar estos tres periodos por separado e integrarlos únicamente a nivel del potencial de recuperación general. En todos estos procesos, se consultó el AED como herramienta para explorar cómo podría formarse un índice general razonable. Antes de llegar a eso, es necesario comentar los puntos críticos del estudio original. Estos pueden condensarse en tres temas, a saber

- ponderación objetiva frente a ponderación subjetiva,
- desarrollo y la influencia relativa del tiempo, y
- el equilibrio entre los factores de vulnerabilidad y riesgo y los recursos.

A continuación profundizaremos en estos temas.

5.5.7.1 Ponderación de los acontecimientos vitales

Las personas perciben el mundo y lo procesan de forma diferente en función de su entorno físico-social-biográfico. ¿Cómo ponderar entonces acontecimientos que objetivamente tuvieron lugar vistos desde fuera, pero cuya representación subjetiva se nos escapa? Para aclarar este problema complejo y difícilmente resoluble, nos preguntamos, a modo de ejemplo si la pérdida de un peluche querido a los 5 años es tan grave como la pérdida de la madre. Ésa es probablemente una pregunta más fácil, pero ¿y ésta? si la pérdida de un padre es tan grave como una exclusión social posterior o una discapacidad física o sufrir violencia física y abusos sexuales? ¿Es esto peor a los 5, 7 o 9 años? Malo es malo, no hay graduaciones finas como la edad, aunque hay que tener en cuenta que la edad desempeña naturalmente un papel en el procesamiento, pero esto está mediado por muchos factores. Los acontecimientos traumáticos antes de la capacidad de verbalizar y antes de la formación del yo son claramente más graves y más difíciles de procesar que los acontecimientos en los que ya existe un yo integrado y se dan las capacidades verbales.

Una razón principal es que es difícil acceder a ellos porque los recuerdos se encuentran en el ámbito no lingüístico. Por tanto, la cuestión de la ponderación no puede responderse ni de forma arrolladora ni objetiva, y tan poco por la frecuencia de aparición de tales acontecimientos vitales críticos (véase también sobre el papel de la resiliencia, Hildenbrand, 2006b). En el mejor de los casos, una reconstrucción de las consecuencias causa-efecto puede conducir a una evaluación. Ahí reside la verdadera razón por la que aquí se ponderan por igual las vulnerabilidades y los recursos. En cuanto a los recursos podría hacerse la misma pregunta: "¿Cuál tiene más peso?". Y como en el caso de las vulnerabilidades esta pregunta no puede responderse de forma radical u objetiva: ¿Es la terapia, el apoyo en un grupo, la atención recibida en el contexto general o simplemente el tiempo? ¿Es imperativo ganar comprensión sobre el propio pensar, sentir y actuar? Es de suponer que no se trata de ninguno de estos factores por sí solo, sino de un conjunto complejo, cuyo verdadero significado es muy probable que se pierda al intentar analizar las puntuaciones realizadas y separarlas del flujo del tiempo, aunque pueda verbalizarse con éxito una ponderación. Por supuesto, podríamos preguntar a las propias personas, pero ¿cómo compararemos entonces las respuestas? ¿Cómo podemos estar seguros de que el peso subjetivo en el momento actual y tantos años después de la terapia corresponde exactamente a lo que era relevante en la terapia o en el momento de la experiencia del trauma o de la apropiación o descubrimiento de un recurso? Y aún es más difícil comparar esto adecuadamente entre personas para introducir una ponderación por persona o entre personas y comparar los resultados. Sería concebible clasificar las vulnerabilidades o los recursos por persona y entre ellas y

asignar puntos de ponderación individualmente. Esto no se hizo en el presente estudio, ya que los factores sólo se obtuvieron mediante el análisis reconstructivo y después de las entrevistas – pero sería una posibilidad a explorar en futuros estudios. Del mismo modo, podría debatirse la cuestión de *lo suficiente y lo necesario*, es decir, ¿qué acontecimientos vitales causan *necesariamente* el trauma o qué acontecimientos causan *necesariamente* la resiliencia y los recursos? ¿Existe algún acontecimiento que obligue a algo y que tenga muchas probabilidades de provocar acontecimientos específicos y concretos, tanto en sentido positivo como negativo? ¿Qué acontecimientos pertenecen al ámbito de la necesidad, pero no son suficientes? ¿Tienen algún sentido estas preguntas?

Para simplificar, los riesgos y recursos verificables se siguen ponderando por igual y se suman. Por otro lado, se introducen pesos justificables dentro de las categorías. Por ejemplo, el consumo de drogas duras (como la cocaína o la heroína) durante la terapia se pondera más que el de drogas blandas (como el cannabis). Y en el período de catamnesis, un círculo de amigos más amplio que se declara de apoyo recibe un peso mayor que un círculo de amigos más pequeño o que ningún círculo de amigos. Se sigue el mismo procedimiento con otras categorías. Este tipo de diferenciación parece legítima y no invalida las explicaciones anteriores.

5.5.7.2 Trayectorias de desarrollo

Si nos preguntamos por la ponderación de los acontecimientos, también nos preguntamos por los procesos de desarrollo. ¿Cómo cambian los pesos, qué influencia tiene el tiempo y cuál es la influencia de los acontecimientos anteriores sobre los posteriores? ¿Es tan alta la influencia de la pérdida del peluche antes mencionado a la edad de 5 años que sigue teniendo un efecto duradero en la vida adulta? Sin embargo, en lugar de una pregunta tan casi incontestable, el problema puede reformularse en términos más manejables ¿Cuáles fueron las consecuencias de la pérdida? ¿Cómo se actuó y sintió de forma diferente? ¿Qué se evitó, qué se reforzó? Cambiemos de nuevo de tema y pasemos a la pérdida de un padre. La pérdida de un progenitor implica muchas consecuencias que pueden durar muchos años y si no se trata puede incluso transmitirse a la siguiente generación como un complejo de problemas que sigue vigente. La herencia de acontecimientos sociales traumáticos y complejos de problemas a lo largo de generaciones es un fenómeno bien conocido en las psicoterapias. Bajo la palabra clave "generación de hijos y nietos de guerra" (Segunda Guerra Mundial) hay un gran número de libros especializados (Bode, 2005, 2008; Lorenz, 2005; Ustorf, 2008; Radebold, Bohleber & Zinnecker, 2008), que se ocupan de este ámbito y cómo los complejos de problemas se trasladan de una generación a otra. El tema de la guerra es, por desgracia, especialmente llamativo y los fenómenos mencionados son "relativamente" fáciles de estudiar y rastrear con los datos disponibles.

Si ahora se examina catamnésicamente lo que hacen realmente las personas - por ejemplo, trabajo, actividades de ocio, relaciones, conductas de riesgo, tendencia a cometer actos delictivos, implicación social, deportes, etc. – y se trata de compararlo con las vulnerabilidades previas y potenciales no tratamos de introducir una ponderación numérica comparativa, sino que nos ceñimos a la simple suma. El principio rector es la idea de que cuanto más diversas y heterogéneas sean las actividades humanas, más sanas y recuperadas estarán estas personas. Los problemas surgen cuando hay restricción y concentración en unas pocas áreas de la vida. Si la gente intenta esto y aquello a pesar de las condiciones desfavorables – contamos también el intento, no sólo el éxito – lo tomamos como un signo de relativa salud y relativas aspiraciones de autonomía. Si las personas hablan ni de objetivos ni de actividades, tendemos a interpretarlo como estancamiento o incluso regresión. En términos numéricos, la ausencia de menciones significa que no se están añadiendo recursos, lo que equivale a que el statu quo anterior (= pre a la terapia) no ha cambiado o ha cambiado muy poco.

Este enfoque resuelve el problema de crear un índice de recuperación con un umbral crítico (véase más arriba), que obviamente ha fracasado. Por el contrario la evolución puede observarse, resumirse y compararse entre los antiguos clientes. Dado que cada uno configura su propia vida en función de sus antecedentes biográficos, no evaluamos de forma diferente si alguien hace esto o aquello. En cambio, valoramos de forma diferente lo que parece posible en el contexto de las condiciones previas. Beat Kaiser, un cliente con un coeficiente intelectual estimado en el rango de superdotados, que, sin tener el título de bachillerato, consiguió ser admitido en una terapia, consiguió ser admitido en una escuela técnica superior en el primer intento. Tras estudió trabajo social y se graduó. Esto lo valoramos justificadamente más que una pura

formación o unas simples prácticas dentro de la categoría de "formación". Sin embargo, no se detiene ahí. Desde una perspectiva de desarrollo, por ejemplo, sigue siendo cierto que Beat Kaiser solía tender a abusar de su intelecto – que le permitió estudiar y graduarse – para saltarse las normas y a entregarse a una desafortunada tendencia al cinismo. Ahora bien, el desarrollo en el plano operativo del análisis de datos no significa negar esta vulnerabilidad anterior, es decir, el abuso de su intelecto, exclusión, etc., sino que simplemente añadimos las nuevas habilidades en el uso del intelecto. Así, tanto el intelecto que sustenta el cinismo como el intelecto que hace posible el estudio entran en el potencial de recuperación en pie de igualdad: uno como vulnerabilidad, el otro como recurso.

Así, en lugar de comparar compulsivamente vulnerabilidades con recursos, operacionalizamos como perspectivas diferentes de un mismo fenómeno: la autonomía humana. La ausencia de nuevas vulnerabilidades y riesgos puede entenderse ciertamente como un recurso, del mismo modo que la falta de adquisición de nuevas competencias (de cualquier naturaleza) puede, a la inversa, dar indicios de vulnerabilidades. En la conclusión del estudio catamnesis, Gürtler, Studer y Scholz (2012), al integrar los resultados en un modelo de etapas relativo a la salida de la adicción, llegan a la etapa superior relevante de "transformar las vulnerabilidades en recursos", lo que puede ser todo un arte. Y esto es exactamente de lo que se trata.

A partir de la cuestión de la relación entre vulnerabilidades y recursos, se puede formular el último punto: el equilibrio.

5.5.7.3 Equilibrio entre recursos y vulnerabilidades

Tras la ponderación y el desarrollo, nos preguntamos finalmente si las vulnerabilidades pueden ser debilitadas o incluso anuladas directamente por los recursos – dependiendo del caso individual, por supuesto. Pero, ¿cómo funciona esto? ¿Se elimina algo, se añade algo o el fenómeno en cuestión cambia? En última instancia, no sabemos lo que ocurre realmente en una persona. Sin embargo, como se ha señalado en la sección anterior, esto puede eludirse simplemente añadiendo nuevas capacidades al modelo general de vulnerabilidades y recursos. Lo mismo ocurriría si se añadieran nuevas vulnerabilidades, como una recaída en el consumo de drogas, la pérdida del trabajo como posible desencadenante de una crisis, etc. Lo mantenemos deliberadamente simple, es decir, igual ponderación de los factores independientemente del signo, sumatoria y sólo subdividida según los puntos temporales antes de la terapia, durante y después (= período de catamnesis). A partir de esto, se puede formar el cociente de recuperación y éste puede utilizarse como medida relativa de comparación entre los antiguos clientes. La diferencia entre los distintos momentos es importante para obtener una imagen global del desarrollo y, por tanto, de la posible compensación de los problemas existentes a través de los recursos. Si sólo nos fijamos en las secciones, éstas pueden distorsionar la imagen. Así ocurrió en el estudio de Gürtler, Studer y Scholz. La tabla 5.11 muestra, por ejemplo, que Beat Kaiser y Julia Nagel muestran un potencial de recuperación comparable antes de la terapia. "Casualmente tienen el mismo número de recursos y vulnerabilidades, por lo que uno podría inclinarse a pensar que ambos tienen el mismo potencial o los mismos retos por delante. Como muestra el estudio catamnesis muestra, sin embargo, este no es el caso. Esto también puede expresarse a partir de las reconstrucciones de casos de dos antiguos clientes:

Beat Kaiser – El antiguo cliente Beat Kaiser muestra un pronóstico bastante favorable, ya que gracias a su coeficiente intelectual todavía muy alto, siguió estudiando en el sector social después de dejar la terapia y aprobó el examen de acceso con nota. También consiguió graduarse, como ya se ha dicho. Y consiguió desarrollar empatía en sí mismo y, por lo tanto, redefinir comportamientos social desfavorables y aprender gradualmente nuevas formas de actuar. Esto significaba que más recursos eran de esperar, de modo que en el momento del examen, su pronóstico era más favorable de lo que cabía esperar del potencial de recuperación de acontecimientos anteriores. Beat Kaiser también informó sobre actividades deportivas como montar en bicicleta los fines de semana, señaló que quería reducir su consumo de tabaco y habló de buenas y constantes relaciones con su familia de origen, que, debido a diversos golpes del destino, todos los miembros de la familia tendían a crecer juntos y no tendían a dividirse. A nivel privado, también habló de una muy fructífera relación de pareja (a distancia) con una mujer sorda que por su especial forma de ser le ayudó a deshacerse cada vez más de su cinismo al apartarse y ya no podía oír nada cuando él "repartía" verbalmente comentarios cínicos. Al mismo tiempo, sin embargo, ella no

tenía una actitud negativa hacia él, ya que no escuchaba lo que decía. Gracias a unas prácticas en el sector social desarrolló empatía hacia sí mismo cuando trabajaba con personas discapacitadas y poco a poco pudo dejar de centrarse en el *trabajo con adictos*.

Julia Nagel – La cliente Julia Nagel, por ejemplo, era diferente. En el momento del estudio tenía el mismo nivel de integración antes de la terapia (véase el cuadro 5.11) como Beat Kaiser, pero su pronóstico era menos favorable, ya que no era de esperar un aumento de sus recursos en la misma medida que en el caso de Beat Kaiser. Así pues, aunque ciertamente existían relaciones sociales constantes, pero las perspectivas de un contrato de trabajo fijo eran limitadas a pesar de su formación y era más probable que realizara una especie de prácticas de larga duración. La vida cotidiana se parecía a una especie de *empleo en un entorno protegido*, lo que correspondía a un menor aumento de la autonomía. Sus contactos familiares eran escasos y probablemente el consumo de drogas era mucho mayor en el momento de las entrevistas, aunque se tratara de drogas más ligeras como el cannabis. No había indicios de consumo de drogas duras o incluso de delitos relacionados con las drogas.

Desde un punto de vista cualitativo, cabía esperar diferentes cursos futuros para estos dos antiguos clientes. Esto incluía un aumento más fuerte de la recuperación para Beat Kaiser y una curva mucho más agravada para Julia Nagel. Cabía esperar resultados comparables para los otros clientes. Para resolver este problema, se llevó a cabo un reanálisis ampliado de los datos para el presente libro. Se crearon tres tablas: antes, durante y después de la terapia, y las vulnerabilidades y los recursos que pudieron reconstruirse a partir de la entrevista y los documentos del caso. Sólo se asignaron ponderaciones dentro de las categorías (véase más arriba), no entre categorías. A su vez, la tabla 5.12 resume las estadísticas de síntesis y el potencial de recuperación, agregados para los distintos periodos. Los puntos poco claros se aclararon mediante la introducción de una categoría externa "suposiciones del entrevistador" frente a "opinión subjetiva / autoinformes del entrevistado". En cada caso se asumió el caso extremo, pero tanto para los recursos como para las vulnerabilidades, de modo que a nivel relativo el error se minimizara en la medida de lo posible.

La tabla 5.12 muestra de forma impresionante, en contraste con el cuadro 5.11, que el orden y los valores respectivos ahora difieren mucho entre los individuos. Las tablas sólo contienen sumas y cocientes, para que podamos hacerlas en una hoja de cálculo o en nuestra cabeza y no en R, ya que trabajar directamente en tablas tenía la ventaja de poder hacer anotaciones y notas justo al lado de los números. De este modo, las ponderaciones y la selección de categorías relevantes podían hacerse flexibles. Por supuesto, esto sería igual de posible en R. Nuestro objetivo aquí es mostrar que hay diferentes maneras de realizar el análisis.

Debido a las ponderaciones introducidas dentro de las clases, la supuesta igualdad de potencial de Beat Kaiser y Julia Nagel ha cambiado y ahora refleja una clara ventaja para Beat Kaiser, que en este caso se debe a su intelecto y a su mayor apoyo social. Si nos fijamos en el período catamnésico posterior a la terapia, Natalie Lang y Beat Kaiser muestran los mayores y más variados desarrollos a nivel privado y personal, mientras que los otros tres antiguos clientes muestran cambios significativamente menores. Sin embargo, sería erróneo concluir que los otros clientes no han cambiado a mejor, sino que sólo reflejan cambios anteriores al punto de partida de las condiciones difíciles. Creamos algunos índices adicionales (véase la tabla 5.12) y los denominamos *factores de impacto*, que relacionan entre sí el potencial de recuperación de distintos periodos de tiempo y modelan una simple medición del cambio intraindividual.

La aplicación de los índices arroja un panorama muy distinto del esperado. Esto muestra el *cambio relativo*, es decir, lo que las personas han hecho y ganado con respecto a sus condiciones (iniciales). Si una persona tiene unas condiciones iniciales difíciles, es necesario un desarrollo personal, social, emocional, etc. mucho más largo y difícil con mayores con mayores obstáculos. En *términos absolutos*, puede haber incluso ciertos límites superiores que la persona no podrá superar durante su vida. Pero en *términos relativos* – como *cambio intra-individual* – puede lograr más que otra persona que, superficialmente, parece haber conseguido "más". Si observamos ahora este factor de impacto, resulta evidente que el mayor impulso intraindividual lo dio el cliente Samuel Gabler - el único que recibe una pensión del 100% IV y le gusta presentarse con "tengo un golpe en la cabeza"(!). Tampoco sorprende que luego venga Beat Kaiser y no Natalie Nagel, ya que el progreso absoluto de Beat Kaiser también es muy alto y tiene el efecto correspondiente. A partir de las reconstrucciones de casos en Gürtler, Studer y Scholz (2012), esta nueva e inesperada clasificación podría estar probablemente bien justificada en términos de contenido, que no tratamos más adelante.

Esto demuestra el beneficio de AED en el descubrimiento de nuevas estructuras y – aquí – en honrar los cambios individuales.

Tabla 5.12: *Estudio de catamnesis, Gürtler, Studer y Scholz (2012, ampliación del potencial de recuperación)*

Potencial de Recuperación (PR)	Sentido	Lang	Kaiser	Nagel	Gabler	Widmer
Pre	Situación inicial	0.93	0.54	0.38	0.19	0.46
durante	durante la terapía	0.67	0.14	0.11	0.33	0.39
Post	período de catamnesis	11.30	9.53	2.71	6.00	4.80
Pre + durante	situación después de la terapía	1.60	0.68	0.49	0.52	0.85
durante + Post	sit. después de la terapía y período de catamnesis	11.97	9.68	2.82	6.33	5.19
Pre + durante + Post	estado absoluto	12.90	10.21	3.20	6.52	5.65
Factor de Impacto$_1$ (PR_{pdp}/PR_{p-})	(Pre + durante + Post)/Pre	13.89	18.97	8.32	34.78	12.24
Factor de Impacto$_2$ (PR_{-dp}/PR_{p-})	(durante + Post)/Pre	12.87	17.91	7.42	33.32	11.28
Factor de Impacto$_3$ (PR_{-p}/PR_{pd-})	Post/(Pre + durante)	7.06	14.01	5.53	11.54	5.65

5.5.7.4 Contribución de los AED

¿Cuál es la contribución del AED? En este caso, puede afirmarse que el AED puso de manifiesto contrastes y planteó preguntas que posteriormente condujeron a un nuevo análisis de los datos desde nuevas perspectivas. El resultado es ahora un desarrollo de los potenciales de recuperación que puede relacionarse de forma mucho más concluyente con los análisis puramente cualitativos de Gürtler, Studer y Scholz (2012). Se eliminó la escala original del potencial de recuperación y, además, se introdujeron ponderaciones exploratorias dentro de las categorías, pero no entre ellas. A partir de ahí, se pudo formular un nivel "absoluto" alcanzado y una ganancia intraindividual relativa por cliente. Los resultados sugieren que tiene sentido mantener esta distinción a la hora de hacer previsiones.

La cuestión de la ponderación puede diferenciarse aún más en el sentido de que un AED adicional podría mostrar si una mayor ponderación según el aspecto temporal puede ser útil o tiene el efecto contrario. Por ejemplo, el pasado próximo (= periodo de catamnesis) se caracteriza posiblemente por un aumento más significativo de los recursos o vulnerabilidades, y los factores anteriores siguen teniendo efecto, pero son menos significativos. A la inversa, los acontecimientos en fases sensibles (tempranas) del desarrollo pueden tener un impacto especialmente fuerte a largo plazo, para bien o para mal. El aspecto de la "transformación de vulnerabilidades en recursos" (Gürtler, Studer y Scholz, 2012, capítulo 10.3) también podría examinarse para una ponderación correspondiente o una reestructuración fundamental de los conceptos asociados. Parece obvio para el ámbito del tiempo que las ponderaciones no deben modelarse de forma lineal, sino más bien logarítmica o al menos no lineal.

Por tanto, la AED ha cumplido aquí la tarea de cribar el material informativo disponible y influencias significativas como los periodos de tiempo, la valencia de los acontecimientos, la evolución intraindividual, etc., así como plantear cuestiones sobre la estructuración y las interrelaciones de los factores implicados. Con estos resultados, la reconstrucción de un caso concreto puede servir para contrastar aclarando las discrepancias del caso y, por tanto, el aspecto de los métodos mixtos ya está disponible, que la AED promueve explícitamente aquí. A este respecto, el AED ofrece incluso soluciones directamente y no se limita a cuestionar el procedimiento anterior, cosa que también se ha hecho. En el curso de ello, surge la tarea de que los investigadores revisen sus propias convicciones sobre subyacentes y distanciarse de viejos supuestos.

5.5.7.5 Ampliación de AED en la dirección de la estadística bayesiana

La discusión muestra que un potencial de recuperación cuantitativo debe ser un constructo multidimensional. Así, además de las vulnerabilidades y los recursos existentes, podrían añadirse – no sólo de forma exploratoria – dos o tres variantes más, como los recursos potenciales y las vulnerabilidades potenciales, así como la dimensión temporal (véase el cuadro 5.12). Estas adiciones ampliarían el modelo existente a un abanico más amplio de posibilidades. No se trata sólo de acontecimientos vitales que han ocurrido realmente, sino también de la evaluación de acontecimientos que no han ocurrido, cada uno de los cuales puede tener una probabilidad condicional o incondicional. De este modo, podrían establecerse conexiones entre los acontecimientos que han ocurrido y los acontecimientos asociados sobre la base de vínculos causa-efecto. Se trata de la idea de las cadenas de Markov: ¿cómo se relacionan las probabilidades de los acontecimientos asociados? Esto está en consonancia con la investigación sobre los acontecimientos vitales críticos, de modo que los acontecimientos no están aislados, sino que se sitúan en un conjunto de relaciones estrechas y entrelazadas. En un sentido más amplio, esto corresponde a un modelo de vulnerabilidad-estrés--afrontamiento en el que el resultado no es fijo, sino que surge dinámicamente debido a la secuencia y el procesamiento subjetivo de los respectivos acontecimientos. La diferenciación de los niveles absoluto y el nivel comparativamente alcanzado interindividual, por un lado, y los cambios intraindividuales en el trasfondo de las propias condiciones iniciales, por otro, abren interesantes aspectos de modelización.

Una ampliación fundamental de este modelo consiste en incluir los acontecimientos traumáticos de generaciones anteriores, lo que facilita la comprensión del curso de algunas biografías y que también se utilizó en el presente estudio de catamnesis (Studer, 1998; Gürtler, Studer & Scholz, 2012). Especialmente en el caso de traumas relacionados con la guerra, la herencia social de complejos de problemas no resueltos es particularmente impresionante (entre otros, Stachowske, 2002). Sin embargo, esta extensión no solo se aplica a la experiencia de las guerras, sino en general a los problemas no resueltos que se transmiten de forma no intencionada e inconsciente a la siguiente generación y, por tanto, se "heredan". transmitidos a la siguiente generación y, por tanto, "heredados". En el plano biológico, la epigenética es una prometedora para explicar técnicamente la herencia de experiencias entre generaciones.

Los posibles acontecimientos mencionados se explican por sí mismos, incluida una posible ponderación basada en todos los acontecimientos. En el caso concreto, nos preguntaríamos: "Si una persona consume drogas duras durante un largo periodo de tiempo, ¿qué experiencias previas en la biografía son probables y cuáles no?" y "Si la persona se encuentra actualmente en este punto, ¿qué podemos o debemos esperar para el futuro?" o "¿Qué acontecimientos a través de las generaciones hacen que la adicción emerja como un problema importante en una generación?". De aquí podría derivarse una red de factores de influencia concretos, que potencialmente podrían tener un impacto en un caso específico. Por ejemplo, es significativo si – y por qué – *no se produce* un desarrollo potencial y esperado y cuál se manifiesta en su lugar. A partir de este programa de contraste, puede derivarse un pronóstico basado en casos concretos. Estas derivaciones pueden utilizarse posteriormente de forma cuantitativa a partir de los datos cualitativos. Esto equivale aproximadamente a utilizar el conocimiento previo y el conocimiento contextual y, de este modo, estimar una distribución a priori de la estadística de Bayes para realizar predicciones (es decir, probabilidades posteriores) antes de disponer de la información real, así como para realizar una determinación posterior de los acontecimientos.

El estudio de caso debe demostrar que no se trata sólo de utilizar técnicas de transformación y descripción cuantitativa de datos para obtener conocimientos, como parece tras leer la mayoría de los libros sobre AED. Más bien, las cuestiones de investigación van definitivamente mucho más allá y es necesario incluir datos que pueden no estar disponibles directamente, sino sólo indirectamente como conocimiento experto o incluso más inespecíficamente como potencial, es decir, como una evolución favorable o desfavorable típicamente esperada. A estos potenciales se les puede asignar sin dificultad una probabilidad justificable en el marco de la estadística de Bayes (véase el capítulo 6). También es concebible no asignar ninguna probabilidad, sino intentar reconstruir estructuras (similitudes, diferencias) en los datos disponibles. Veríamos la estadística bayesiana como un suplemento o instrumento, pero no como un sustituto en el sentido de reintroducir un método orientado estrictamente según la lógica de pruebas. La atención se centra mucho más en la verosimilitud (Pólya, 1954a, 1954b) y la robustez (Tukey, 1977) en nuestras actividades de investigación.

Antes hemos hablado repetidamente de ponderaciones, pero en última instancia no son más que la asignación de una probabilidad en relación con otros sucesos. Sin embargo, independientemente de esto, podemos asignar ponderaciones (probabilidades) sin la mencionada red de factores influyentes y esto depende de cuánto sepamos sobre un tema. Cuanto menos sabemos, más tenemos que confiar en la intuición o el juicio de los expertos. Cuanta más información tengamos procedente de estudios y otras fuentes, menos influirán la intuición experta y el juicio subjetivo. Esta lógica de asignación de probabilidades a los sucesos incluso sin recuento previo dentro de la estadística de Bayes permite combinar de forma flexible el conocimiento cualitativo con algoritmos objetivos. Tales procesos serán el tema central del próximo capítulo. Con esto caminamos en el nivel de superposición cualitativo y cuantitativo en pie de igualdad y dentro del cuantitativo entre AED y la estadística bayesiana. Llegados a este punto, sería posible calcular una distribución de probabilidad posterior y utilizar dicha información a priori (por ejemplo, de otros estudios, bibliografía profesional, entrevistas a expertos, etc.) para obtener un potencial de recuperación más exhaustivo PR+ en contraste con el anterior potencial de recuperación simple RP para cada uno de los casos mencionados incluyendo límites de tolerancia al alza y a la baja. Los límites de tolerancia permiten formular cuantitativamente el aspecto de la robustez.

A nivel de los datos originales, se dispone de información en forma de frecuencias o datos binarios, de modo que podría visualizarse la proximidad y la distancia entre los casos o recurrirse a la agrupación o a un procedimiento similar con técnicas AED comparable al estudio de caso anterior sobre el liderazgo (véase el capítulo 5.5.5). Un enfoque bayesiano no contradice categorizar los datos visualmente según diferentes aspectos. Esto conduce a posibles ponderaciones e interconexiones, como qué factores de entrada están interrelacionados y cuáles no (véase el análisis de prototipos en el estudio de caso sobre liderazgo, s. Fig. 5.30, p.425) y cómo afecta esto a la cercanía y la distancia entre las personas estudiadas a partir de las vulnerabilidades y los recursos reconstruidos.

> Tarea 5.4: Visualización por proximidad y distancia
>
> Los lectores interesados pueden intentar su propia visualización utilizando sus propios datos de casos o codificación. Hay suficientes sugerencias en el estudio de caso sobre el liderazgo en el sistema educativo español en el capítulo 5.5.5. Asimismo, puede intentar realizar ponderaciones basadas en el contenido de la información entrante y comparar los resultados con los resultados no ponderados. El objetivo sería comprender que la forma en que se prepara la información determina el aspecto posterior de los resultados. Esto se aplica tanto a la recogida de datos como a todos los pasos intermedios.

Los resultados y el enfoque en profundidad que se expone a continuación podrían enriquecer los conocimientos sobre las trayectorias catamnésicas, no sólo para la drogadicción, sino también para una amplia variedad de desarrollo humano. El uso en el contexto de las intervenciones se presta independientemente de si el desarrollo es patológico (por ejemplo, la adicción) o no, un desarrollo normal (por ejemplo, en la escuela) o incluso un desarrollo hacia el máximo rendimiento (por ejemplo, el deporte, la formación de líderes, etc.). El conjunto lógico básico sigue siendo el mismo. En este sentido, las personas y las instituciones pueden utilizar esta información para intervenciones adecuadas al caso y la situación, y en contextos de aplicación muy diferentes.

5.6 Debate sobre el AED

Las observaciones anteriores sobre el AED perseguían varios objetivos - además de concienciar sobre las soluciones creativas nos gustaría mencionar otra idea basada en Tukey. Aunque la estadística inferencial intenta extraer conclusiones de las muestras a las poblaciones con un cierto nivel de error, desde cierto punto de vista no vemos tanta diferencia con la estadística puramente descriptiva con ayuda de técnicas creativas,

es decir, AED. La estadística inferencial surge del hecho de que un procedimiento "algorítmico-descriptivo" se añade un umbral crítico, de modo que, por ejemplo, puedan tomarse decisiones de prueba. Esto significa cualitativamente que a un algoritmo y a sus resultados concretos se les asigna un significado que no es inherente a los algoritmos ni a los datos, sino que procede de convenciones y hábitos o, en el mejor de los casos, de consideraciones humanas bien fundadas. Sin embargo, la práctica demuestra – tales umbrales proceden simplemente de convenciones y pautas y no se eligen adecuadamente para el caso. Volvamos a Bateson (1985), que define la información como una diferencia que marca realmente una diferencia. Esto debe interpretarse en el sentido de que la información (diferencias) reciben y contienen significado. Entonces, las pruebas de significación convencionales sólo contienen un significado trivial (información), a saber, el de la convención y no el vínculo crítico adaptado al caso. ¿Cuál es entonces la ganancia de conocimiento y ¿es pronunciada?

Siguiendo estos argumentos, la estadística inferencial clásica puede entenderse en el sentido de que, en función de las variables presentadas (p-valores, tamaño de la muestra, potencia, etc.), sólo se observa el comportamiento de cómo cambian estas variables ante cambios concretos y nominativos (muestras empíricas, nuevas muestras, instrumentos de medida cambiados con nueva potencia, supuestos de error cambiados, modelo cambiado con predictores diferentes, etc.). Esto no significa prescindir de las inferencias, que a menudo son necesarias por razones prácticas. Pero desde esta perspectiva, todos estos modelos pueden entenderse como variables dinámicas que describen un proceso complejo. El ritual (cero) (véase el capítulo 4.3.8) sobre la cuestión aislada de la significación estadística puede así pasar a un segundo plano con tranquilidad. En última instancia, los valores *p* son simples transformaciones no lineales de los datos (véase el capítulo 4.3.9.1) a lo largo de un conjunto definido de algoritmos. Puesto que son una propiedad de los datos, a saber la probabilidad de éstos bajo la validez de la hipótesis nula, se trata de una transformación de los datos y, por tanto, no de un sesgo en el sentido convencional, pero tampoco de una mejora o una ganancia real de conocimiento. En última instancia, los algoritmos estadísticos algoritmos se limitan a reducir la complejidad de muchos datos paralelos, que ocultan las tendencias centrales en su crudeza. La importancia de los valores *p* sólo aparece cuando se añade la cantidad teórica llamada significación. Sin embargo, si ésta no está bien fundamentada, cabe preguntarse con razón si la prueba de significación no produce entonces una visión sesgada. Sin embargo, el sesgo reside entonces en la falta de fundamentación de carácter sustantivo y no en el hecho de que los datos se hayan procesado de forma reducida. Esta última es una forma legítima de ver los datos, no otra cosa hemos practicado aquí en el capítulo sobre AED sin parar. Los datos se transformaron y se miraron desde distintos ángulos. Si tuviéramos criterios claros a lo largo de cada una de estas perspectivas – por ejemplo, afirmaciones sobre distancias mínimas o máximas en el estudio de caso sobre el liderazgo -- sería fácil introducir la noción de significación y decisión de prueba o confirmación y prueba por la puerta de atrás. Nos hemos abstenido deliberadamente de hacerlo y nos hemos mantenido en el nivel incierto de la perspectiva. En nuestra opinión, esto es mucho más exigente, ya que requiere la interpretación y la búsqueda de significado, que la prueba de significación en su procesualidad algorítmica no requiere e ignora.

Por lo tanto, creemos que es mejor poner la descripción de la *complejidad* en primer plano para poder tomar decisiones razonables. Para ello, es necesario decidir qué modelo tiene sentido y cuál no, en función de las distintas variables influyentes. Cuando se trata del comportamiento en la toma de decisiones, necesitamos naturalmente un criterio bien fundado y un proceso que regule en qué condiciones y cómo se toman las decisiones y que también sopese cuidadosamente las respectivas consecuencias prácticas. Del mismo modo, nunca debemos confundir la exploración con la confirmación. En el lado de las dificultades, por supuesto, está el hecho de que tal procedimiento no funciona de forma generalizada y no siempre, sino que siempre requiere ajustes inesperados. Así es como podría funcionar la realidad.

Lo interesante es precisamente la *zona gris* en la que los datos parecen indiferentes, ni realmente "aleatorios" ni realmente "sistemáticos." Por tanto, sugerimos que en estas situaciones se utilice el pensamiento cualitativo en combinación con técnicas del AED para basar las decisiones en múltiples perspectivas. El AED hace aquí una contribución creativa e innovadora, o quizá la contribución más creativa e innovadora.

Capítulo 6
El Razonamiento plausible – La Estadística de Bayes

> »One sees, from this Essay, that the theory of probabilities is basically just common sense reduced to calculus; it makes one appreciate with exactness that which accurate minds feel with a sort of instinct, often without being able to account for it.«
>
> Introduction to *Théorie Analytique des Probabilités*
> Pierre-Simon Laplace, 1812

6.1 Objetivos del capítulo

El capítulo sobre la estadística de Bayes trata de un debate resumido sobre epistemología y actitud, instrumentos y medios, criterios de bienes y ventajas e inconvenientes de la estadística de Bayes, incluyendo algunos estudios de casos con R para inspirar y estimular el pensamiento. La atención se centra en el tipo de pensamiento que guía la estadística de Bayes. No es ni un texto de enseñanza ni un manual. Para ambos, véanse los libros de Jaynes (2003), Gelman, Carlin, Stern y Rubin (2003), Gelman y Hill (2007), Bolstad (2007), Gregory (2006), Kruschke (2015b) y McElreath (2015), por nombrar solo algunos. No nos es posible siquiera acercarnos a reproducir la amplitud y profundidad de los debates que han tenido lugar, y mucho menos detallar todos los escollos, consejos y trucos, etc., además de las matemáticas necesarias. El objetivo sigue siendo comprender los métodos mixtos, para que los lectores interesados en Bayes se informen más.

En cuanto a R - como de costumbre, R tiene su propia página general muy rica sobre Bayes (CRAN, 2019a). Hay algunos paquetes de R que permiten escribir código para OpenBUGS (2019), JAGS (2019) y Stan (2019b) a través de R para activar y examinar los resultados. BUGS, JAGS y Stan son los programas de simulación MCMC más comunes y potentes actualmente disponibles para la estadística bayesiana. OpenBUGS, BUGS y JAGS son muestreadores de Gibbs (s. cap. 6.13.1.2), mientras que Stan es un representante del muestreo HMC (s. cap. 6.13.1.3). El código se puede ejecutar con `rbugs` o `BRugs` (para OpenBUGS), `rjags, R2Jags` y `runjags` (para JAGS) y `rstan` o `brms` (para Stan). El algoritmo Metropolis-Hastings apenas se utiliza hoy en día (véase el capítulo 6.13.1.1), ya que el muestreo de Gibbs y el muestreo HMC son algoritmos mucho mejores, más eficientes y más robustos. Los paquetes adaptados al aprendizaje bayesiano en el contexto de los libros de texto incluyen `BaM` (Gill, 2007), `Bolstad` (Bolstad, 2007), `LearnBayes` (Albert, 2007) y `rethinking` (McElreath, 2015). Kruschke (2016d) también ofrece una amplia colección de scripts relacionados con su libro. Los análisis posteriores de simulaciones MCMC pueden realizarse con `coda` y `bayesplot`. `BayesianFirstAid` y `BEST` ofrecen variantes de Bayes para pruebas comunes. Los factores de Bayes están disponibles con `BayesFactor`.

6.2 Problemas básicos: incertidumbre, estimación, decisión

Entonces, ¿qué es la estadística bayesiana y por qué parece ofrecer tanto material que aparentemente sigue siendo suficiente para alimentar décadas de disputas entre científicos? Tanto es así que la profesora emérita de filosofía Deborah G. Mayo (2018) dio a su último libro el subtítulo "How to Get Beyond the Statistics Wars", que luego lamentablemente no está a la altura, ya que ella misma sólo defiende una variante particular y no integradora de la estadística. Antes de decir más, intentemos una introducción basada en problemas a la estadística bayesiana en lenguaje cotidiano utilizando varios pequeños casos prácticos, incluidos los diagnósticos médicos, para que podamos preguntarnos después si los médicos y los asesores médicos son capaces de interpretar correctamente sus propias pruebas médicas. Un estudio de Gigerenzer, Horage y Ebert (1998) da fuertes indicios de que no es así, porque el pensamiento bayesiano con probabilidades condicionales aún no está muy extendido (véase también Gigerenzer, 2004a). La estadística bayesiana aporta ideas sobre cómo afrontar las incertidumbres de la vida, siempre relacionadas con las decisiones concretas que conllevan consecuencias reales. Esto es especialmente evidente en las decisiones judiciales. A menudo, los testigos sólo pueden reconstruir acontecimientos y son sensibles a recordar cosas que no sucedieron (Loftus, 1998). Así pues, los juicios dependen de cómo se construya científicamente la argumentación subyacente. Lindley (2000, p.317) lo describe así: „[t]hese aspects concern the trial process, where there is uncertainty about the defendant's guilt, uncertainty that is subsequently tempered by data, in the form of evidence, hopefully to reach a consensus about the guilt." Un acusado es realmente culpable o no. Esta culpabilidad es incierta y se expresa como una probabilidad. Las pruebas y otros datos modifican esta probabilidad, de modo que la probabilidad de *culpabilidad* se contrapone a la de *no culpabilidad*.

Otros usos de la estadística bayesiana son el análisis de riesgos, de modo que cuando se toman decisiones, a cada opción se le asigna lo que se denomina una función de pérdida, que nos dice cuál es el coste de elegir esa opción. Así, en el caso de las sentencias judiciales, habría que considerar qué tiene un efecto peor: ¿un inocente en la cárcel o un culpable en libertad? De forma equivalente, uno podría plantearse si prefiere invertir en bienes inmuebles o en acciones o en fondos. ¿Qué posibilidades hay de que el dinero aumente o, al menos, de que no se produzcan grandes pérdidas? Unos beneficios potencialmente mayores van acompañados de mayores riesgos y, en consecuencia, la estadística de Bayes puede ayudar a cuantificar estas variables abstractas. Además, se pueden ejecutar escenarios óptimos y peores para examinar todo el espacio de posibilidades de las estructuras causa-efecto y extraer conclusiones coherentes y razonables.

No hay que olvidar la ciencia en sí misma, no sólo el trabajo científico y el pensamiento al servicio de una institución como el tribunal o la bolsa, sino el proceso de generar conocimiento, desenmascarar el falso conocimiento como tal, desarrollar métodos de intervención y, en general, explicar el pasado y predecir el futuro (véase el capítulo 1.2). Estadística es un aspecto de la exploración de datos que sirve de apoyo a la ciencia creando, probando, descartando y mejorando modelos. Existen diferentes perspectivas sobre cómo tratar los datos y los modelos:

- Probar los modelos entre sí o ampliar los modelos y probar los componentes entre sí (en general, factores de Bayes, véase el capítulo 6.8.1). Los factores de Bayes no corresponden a un enfoque bayesiano completo, ya que sólo se trata de actualizar las expectativas y no las probabilidades posteriores de los parámetros.
- Integrar modelos y mantener o aumentar la complejidad (enfoque bayesiano completo).

La incertidumbre propiamente dicha también puede expresarse de dos formas (Lindley, 2000, p.300):

- Confianza – Probabilidad de que un intervalo (= rango de valores) contenga una cantidad buscada, el parámetro ,.
- Probabilidad – Probabilidad de que una cantidad buscada, el parámetro , se encuentre en un determinado intervalo (= rango de valores).

El primer caso es la probabilidad en un intervalo con respecto al parámetro . El segundo caso denota la probabilidad del parámetro sobre un intervalo, es decir, la suposición de un cierto intervalo de valores.

6.3. Estadística de Bayes – una introducción

Bayes es la vida cotidiana y la vida cotidiana es Bayes. Una introducción a Bayes debe empezar por la vida cotidiana, por el sentido común, como defendían Laplace, Pólya o Jaynes (Studer, 1996b). Cuando nos levantamos por la mañana, podemos preguntarnos qué probabilidad hay de que nuestro té ya esté hirviendo, de que nadie nos moleste mientras trabajamos hoy en la oficina o de que no nos encontremos en un atasco de camino a casa.

Si todo el mundo se hace estas preguntas, enseguida queda claro que todos podemos expresar opiniones más o menos inequívocas sobre estos acontecimientos tan concretos y sus probabilidades asociadas. Esto es así aunque no tengamos (no podamos tener) ni idea de los hechos respectivos. Todo el mundo puede decir algo sobre el tiempo, pero ¿quién puede dar información fundada sobre él o incluso predecirlo? En estos casos, cuando no pensamos más en variables de influencia específicas y muy concretas, como nuestro té de la mañana, el tiempo, los compañeros o los coches de camino al trabajo y de vuelta, asumimos probabilidades que pueden asignarse al *tipo incondicional*.

El término incondicional significa simplemente que no consideramos la ocurrencia de sucesos en dependencia directa de la ocurrencia de otro suceso muy concreto, sino de forma casi general, es decir, teniendo en cuenta todas las posibles variables influyentes que pueden tener un efecto. Ninguna de estas posibles influencias recibe un trato preferente a la hora de estimar la probabilidad incondicional de un suceso. De este modo, nos comportamos como si no existiera una realidad concreta delimitada, sino que la realidad sería si todas las posibilidades se dieran paralelamente, aniquilando así por completo cualquier contexto. Esto es como caminar por el espacio de posibilidades potencialmente infinito y luego mirar realmente en todos los rincones para que no se nos pase nada por alto.

Los cálculos elaborados basados en algoritmos MCMC hacen precisamente eso en la estadística de Bayes: exploran todo el espacio de posibilidades de sucesos (parámetros) y lo hacen de la forma más completa y exhaustiva posible. Si además suponemos que en nuestro mundo no hay sucesos verdaderamente incondicionales, entonces las probabilidades incondicionales no son más que la suma de todas las influencias efectivas. Para ello, es irrelevante que conozcamos o no la ocurrencia de todas estas condiciones. En principio, se trata de una cantidad infinita, pero en la práctica esta cantidad se reduce considerablemente. La práctica no conoce cantidades infinitas, sino sobre todo un número manejable de influencias, que por tanto no están en absoluto sujetas a nuestro control.

Si nos centramos ahora en esas influencias concretas que hasta ahora hemos ignorado en vista de todos los acontecimientos posibles, aparecen las *probabilidades condicionales* de los acontecimientos. El término condicional significa ahora que evaluamos nuestros sucesos de entrada (té, tiempo, comportamiento de los compañeros, situación del tráfico) en relación con la existencia (=ocurrencia) de otros sucesos concretos y ya no de forma global, es decir, incondicionalmente y ante el espacio infinito. Estas condiciones pueden ser, por ejemplo, las siguientes:

- ¿En qué medida nuestro té matutino depende de que haya té en la casa? Porque, independientemente de que el agua ya esté hirviendo, la taza de té esté lista, el infusor esté enjuagado y caliente, etc., si todas las latas de té de la casa están vacías, la probabilidad de un té matutino (y no sólo la de un buen té matutino) disminuye hasta exactamente cero. Por cierto, esto no dice nada en absoluto sobre un posible café en el desayuno.
- Podemos mirar el tiempo en relación con la estación o en relación con el tiempo de ayer. Si ayer hizo sol, ¿esperamos que haga sol hoy? ¿Esperamos temperaturas bajo cero en verano o temperaturas estivales superiores a 30° Celsius? ¿Hemos leído y entendido la previsión meteorológica(para la ciudad adecuada) y lo hemos entendido?
- Equivalente para la situación en la oficina: ¿molesté ayer "accidentalmente" a mis compañeros? ¿me perdonan por ello hoy o no y me dejan sentirlo?
- Y, por último, no hay nada más aplicable a la situación de los coches: si es época de vacaciones, esperamos que haya menos coches circulando por las ciudades porque todo el mundo se ha ido de vacaciones. Pero no

concluimos necesariamente que las autopistas estén vacías durante las vacaciones. Puede ocurrir lo contrario, es decir, que se produzcan atascos, retenciones y un aumento de la siniestralidad. Todo ello provoca retrasos considerables.

Técnicamente, en el caso más sencillo, en términos de probabilidades condicionales, hay dos sucesos que son mutuamente dependientes y que, tomados por separado, también tienen cada uno probabilidades incondicionales dado el infinito espacio de posibilidades, ya que la práctica suele limitarse a un conjunto finito de sucesos y entradas. Explicado brevemente con el ejemplo del té de la mañana (=MoTee), cuando p representa la probabilidad:

- *p(MoTee | té disponible)* – La pregunta principal es "¿Cuál es la probabilidad condicional de que haya té matutino si hay té disponible en la casa?". Para indicar probabilidades condicionales, se utiliza un guión vertical "|" para separar el suceso de interés y la condición.

Esta probabilidad condicional requiere para su cálculo al menos las siguientes probabilidades siguiendo la argumentación anterior, a saber

- *p(té disponible | MoTee)* – la probabilidad condicional de que haya té disponible en la casa cuando hay té por la mañana,
- *p(MoTee)* – la probabilidad incondicional de que haya té por la mañana, y
- *p(té disponible)* – la probabilidad incondicional de que haya té en la casa.

Se puede demostrar (véase también el capítulo 6.4.2) que a partir de estas cuatro probabilidades – dos incondicionales y dos condicionales – se puede deducir el *teorema de Bayes* y, en última instancia, toda la estadística de Bayes. Todo lo demás son matemáticas complejas. Así, en el caso más sencillo, dos sucesos y sus probabilidades individuales incondicionales y condicionales inter-relacionadas se ponen en un orden para enumerar matemáticamente todas las cantidades en el marco de una única ecuación. El punto de partida es la regla del producto:

$$p(MoTee \mid Té\ disponible) \cdot p(Té\ disponible) = p(Té\ disponible \mid MoTee) \cdot p(MoTee) \qquad (6.1)$$

Se puede deducir el *teorema de Bayes* de esta ecuación. En caso de se interese de

p(MoTee | Té disponible)

el teorema de Bayes sigue por re-formulación de la ecuación 6.1:

$$p(MoTee \mid Té\ disponible) = p\,\frac{Té\ disponible \mid MoTee) \cdot p(MoTee)}{p(Té\ disponible)} \qquad (6.2)$$

Si hay tres cantidades, se obtiene automáticamente la cuarta. Esto es idéntico tanto para el caso discreto como para el continuo y también se aplica al caso en que intervienen más de dos variables. Entonces "simplemente" se complica. Así, el té de la mañana podría predecirse en principio no sólo por si nuestro té favorito está disponible en casa, sino también por si nos apetece tomar el té esta mañana, si tenemos tiempo para ello, si el agua está hirviendo, etc. De forma equivalente, se pueden modelizar otros ejemplos como el tiempo atmosférico, el comportamiento de los compañeros, la situación del tráfico, etc.

El teorema de Bayes se define de forma abstracta como sigue:

$$\text{Posterior} = \frac{\text{Prior} \cdot \text{Likelihood}}{\text{Total}} \qquad (6.3)$$

Los términos a *Priori* y a *Posteriori* se refieren al estado de la información antes y después de ver los datos empíricos. La *Likelihood* corresponde al término maximizado en estadística clásica y entonces conocido como método de máxima likelihood, que señala una función de densidad de probabilidad, que depende de ciertos parámetros (por ejemplo, μ y σ^2 en el caso de variables aleatorias independientes distribuidas normalmente). La *Likelihood* es la densidad de probabilidad condicional de los datos dados los parámetros; y recuerde, el valor *p* de la estadística clásica es la probabilidad de los datos dada la ocurrencia de la hipótesis nula. Es decir, si se conocen exactamente los parámetros, ¿cuál es la distribución de los datos? El término *Likelihood* se refiere entonces a la "probabilidad" (es decir, lo probable o plausible) de los datos para una combinación dada de datos y parámetros. En sentido estricto, no es una probabilidad, porque la *Likelihood* puede tomar valores superiores a uno, lo que no puede hacer una probabilidad pura. Sin embargo, no hay mejor traducción que "verosimilitud" para Likelihood, por lo que en la práctica es mejor mantener el término Likelihood y no traducirlo. El método de máxima Likelihood se remonta en su difusión a Ronald A. Fisher, pero fue probablemente ya utilizado en trabajos preliminares por Carl Friedrich Gauss (1777-1855) y Pierre-Simon Laplace (1749-1827). El economista irlandés Francis Ysidro Edgeworth (1845-1926) también la derivó en 1908.

En algunos casos, sólo el numerador se calcula mediante el teorema de Bayes, ya que es proporcional a la "*posterior*" y el denominador "sólo" contiene una constante normalizadora que normaliza el numerador al intervalo cero a uno.

$$\text{Posterior} \propto \text{Prior} \cdot \text{Likelihood} \tag{6.4}$$

El cálculo de la *probabilidad total* en el denominador, también llamada *probabilidad marginal* o *evidencia*, puede ser muy complejo en la práctica, ya que se aplica lo siguiente

$$\text{Posterior} = \frac{\text{Prior} \cdot \text{Likelihood}}{\sum(\text{Prior} \cdot \text{Likelihood})} \tag{6.5}$$

Así pues, el espacio de sucesos completo debe calcularse exhaustivamente. Esto se aplica tanto al caso *discreto* (=calcular suma) como para el caso *continuo* (=calcular integral) y contiene tanto la probabilidad de que se produzca el caso como la probabilidad de que se produzca lo contrario. De forma abstracta, el teorema de Bayes se transforma así para los sucesos A y B.

$$p(A \mid B) = \frac{p(A) \cdot p(B \mid A)}{p(B)} \tag{6.6}$$

Utilizando la *ley de la probabilidad total* (derivada en Jaynes, 2003) para el denominador *p(B)* y tras la descomposición en los conjuntos de sucesos para A y B

$$p(A_i \mid B) = \frac{p(A_i) \cdot p(B \mid A_i)}{\sum_{j=1}^{N} p(B \mid A_j) \cdot p(A_j)} \tag{6.7}$$

se aplica a la descomposición del denominador en el caso discreto con dos expresiones

$$p(A \mid B) = \frac{p(A) \cdot p(B \mid A)}{\left[p(B \mid A) \cdot p(A)\right] + \left[p(B \mid \overline{A}) \cdot p(\overline{A})\right]} \tag{6.8}$$

que el denominador contiene tanto las probabilidades de ocurrencia del suceso B bajo la condición A como la probabilidad de ocurrencia de su complemento *no A*.

Como debería ser obvio a estas alturas, no necesitamos contar, como se practica en la estadística clásica. Si carecemos de experiencia con el té de la mañana, por ejemplo, porque hasta ahora hemos sido bebedores

de café, podemos seguir adelante con el teorema de Bayes en cuanto dispongamos de una cantidad mínima de datos o de información cuantificable procedente de otras fuentes. A pesar del consumo de café durante muchos años, "accidentalmente" hemos resuelto firmemente ser bebedores de té a partir de mañana por la mañana (y solamente entonces disponemos de datos empíricos) y podemos especular sobre lo que ocurriría si fuéramos bebedores de té y utilizar esto como conocimiento pevio (la variable "prior"). O podemos tomar ejemplo de otros bebedores de té. Más allá de eso, no hay nada que contar, porque en nuestro ejemplo no hay valores empíricos, es decir, datos, que puedan cubrir las probabilidades de interés anteriores. Tampoco hay necesidad de una muestra enorme, porque, como recordatorio, sólo somos bebedores de té a partir de mañana y somos un individuo, no un grupo enorme e idealmente distribuido normalmente. Así que falta la base de datos necesaria para la estadística clásica. Como sugiere el ejemplo, en el caso extremo de una muestra mínima (= caso individual) y datos mínimos (= una fecha), no se podría decir nada en absoluto ni calcular nada. No ocurre lo mismo con Bayes: es posible calcular con un solo caso y datos mínimos. Esto tiene ventajas que se aprovechan, por ejemplo, en astrofísica, que en muchos casos se ocupa de sucesos individuales. Por supuesto es perfectamente legítimo y coherente calcular con lo que hay disponible. Esto constituye la realidad del statu quo: lo que sabemos y lo que no sabemos. Del mismo modo, está implícito que tomamos probabilidades razonadas como expertos cuando falta la base de datos, es escasa o tiene sentido por otras razones combinar datos con una opinión experta para expresar un conocimiento previo. No se pretende que esto sea la norma, ni siquiera que refleje suposiciones arbitrarias piadosas. Pero Bayes tiene una visión pragmático-práctica de la información disponible. Si falta información, es un hecho. Entonces se puede estimar – primero subjetivamente como experto o profesional y después a partir de datos empíricos, si se dispone de ellos.

Una probabilidad es una estimación abstracta y no un recuento, por lo que puede proceder de muchas fuentes distintas y no se basa únicamente en frecuencias. La Likelihood es el término de la ecuación (véase más arriba) que procesa los datos empíricos. Así, a la larga, los datos recopilados deberían sustituir a las estimaciones pegadizas de los expertos, de modo que sean los datos los que impulsen el teorema de Bayes y no las estimaciones supuestamente "subjetivas". Este punto de la *subjetividad* (entre otros, véanse los capítulos 6.3.1, 6.5.2 y 6.5.2.4) es el *punto crítico* de la estadística de Bayes desde el punto de vista de la estadística clásica, que no reconoce el conocimiento previo y lo rechaza por completo. Sin embargo, desde el punto de vista de Bayes, el conocimiento previo no es simplemente subjetivo u objetivo, sino un ni-ni. El conocimiento previo refleja abstractamente el estado actual de la información en un momento dado. Si no se dispone de datos, no puede utilizarse nada empírico, por lo que las conclusiones se basan en estimaciones de expertos.

Sin embargo, si se dispone de (grandes) cantidades de datos, las conclusiones se basan (exclusivamente) en ellos y, en general, las estimaciones iniciales de los expertos adquieren así un papel cada vez menor y su influencia disminuye en favor de los datos empíricos que determinan la distribución posterior y, por tanto, las estimaciones.

Esto pone de manifiesto que, especialmente con un bajo nivel de información, muestras pequeñas, etc., la estadística bayesiana tiene una ventaja inestimable que la estadística clásica no puede compensar. E incluso con muestras grandes es capaz de *iniciar el aprendizaje a partir de la experiencia* mediante el principio de actualización del teorema de Bayes, lo que no es posible en la estadística clásica debido a la suposición aleatoria de las muestras, que prohíbe simplemente agregar datos. El teorema de Bayes permite un proceso de auto-actualización al convertir el conocimiento posterior en conocimiento previo de acuerdo con determinadas reglas y utilizarlo así. De este modo, el conocimiento posterior en el momento t1 puede utilizarse para el conocimiento previo en el momento t2 con el fin de calcular el conocimiento posterior en el momento t2.

$$\text{Posterior}_{t2} = \frac{\text{Prior}_{t1} \cdot \text{Likelihood}_{t2}}{\text{Total}_{t2}} \qquad (6.9)$$

Esto puede ampliarse a voluntad. El clou de la estadística de Bayes es que el resultado de un análisis se utiliza directamente como conocimiento previo en el siguiente cálculo cuando se añaden nuevos datos.

Más adelante, también se podría calcular directamente sobre todos los datos: el resultado es siempre el mismo (véase un estudio de caso en el capítulo 6.15.2.1). Las muestras no se separan unas de otras, sino que se integran y así se acumulan conocimientos. El resultado, el estado final del conocimiento, debe ser siempre el mismo, ya que la base de datos global es la misma y debe ser un proceso de razonamiento coherente (Cox, 1961; Jaynes, 2003). Entonces, sólo a la luz de nuevos datos se actualiza el estado de error. Esto es muy similar al concepto de *saturación teórica* en la *teoría fundamentada* (véase el capítulo 9.3). Un procedimiento así sería impensable en la estadística clásica. Allí, los conjuntos de datos no están conectados, se consideran realizaciones aleatorias separadas y distintas de los parámetros verdaderos, lo que por supuesto conlleva un problema importante, a saber, dónde establecer el punto de corte en la recopilación de datos (palabra clave "*reglas de parada*", Gelman, Carlin, Stern & Rubin, 2003). En consecuencia, se necesitan metaanálisis para sortear de algún modo el problema. En la estadística bayesiana, esto también desempeña un papel tanto en la inferencia como en la comprobación de modelos (Gelman, 2014b).

6.3.1 Subjetividad: el principal argumento contra Bayes

La inclusión del conocimiento previo nos lleva más o menos directamente de vuelta al controvertido tema de discusión, la *subjetividad*. Abordamos este importante aspecto varias veces desde diferentes perspectivas. Por ejemplo, en los capítulos 6.5.2 y 6.5.2.4, en el contexto de Bayes, y ya en el capítulo 4.3.7, en relación con la estadística clásica. ¿Hasta qué punto es legítimo estimar el conocimiento como experto cuando *no se dispone de datos, o éstos son insuficientes* – ni siquiera hay *datos suficientes*? Este punto se discute de forma bastante controvertida dentro de la estadística bayesiana y existen corrientes de opinión sobre este tema que prácticamente se excluyen mutuamente. Los representantes (Robbins, 1956) de la escuela de pensamiento *Empirical Bayes* (Bayes empírico) intentan evitar cualquier estimación experta y sólo derivan precisamente esta parte del teorema de Bayes directamente de los datos y nunca sobre la base de una estricta lógica de razonamiento cualitativo. Tal procedimiento – si se obtiene el conocimiento previo directamente de los datos empíricos – tiene el efecto de mezclar la exploración y la confirmación en el mismo conjunto de datos. Otros representantes, sin embargo, a saber, los de la *dirección subjetiva* (de Finetti, 1974; Galavotti, 2001), entienden que todo es exclusivamente subjetivo debido al grado de creencia personal, lo cual es igualmente poco esclarecedor, porque la ciencia debería ser al menos intersubjetiva y, para empezar, esto tiene poco que ver con Bayes. En primer lugar, es importante entender en este punto que el conocimiento experto es, en principio, directamente traducible a una cantidad numérica y a un supuesto, a saber, una probabilidad. Esto tiene sentido si no se dispone de datos empíricos previos que se encarguen de esta tarea o si las razones del objeto investigado así lo exigen o sugieren o si ya existe una buena y discutible argumentación transparente.

E incluso si existen datos, en principio aún pueden ser modificados por el conocimiento experto ante argumentos convincentes para expresar conscientemente una determinada expectativa antes de que las tendencias inconscientes se conviertan en profecías autocumplidas. En lugar de tachar y condenar de subjetivo este proceso de recopilación de conocimientos previos, parece más apropiado denominarlo *asunción de responsabilidades*. Esto es cierto cuando el proceso de generación de conocimiento previo es transparente y está bien fundamentado y no es esquemático, sino que está anclado en la materia. Éstos son los criterios de la ciencia: una lógica del razonamiento que pueda ser vista por todos, que esté abierta al discurso crítico y que, en principio, pueda cambiar y lo hace.

La finalidad de estas estimaciones expertas – sin entrar en las matemáticas subyacentes – es que condicionan el espacio de probabilidades (lo acotan, por así decirlo) y proporcionan así una dirección. Esto significa que el resultado de la pregunta "¿Qué probabilidad hay de que esta mañana haya té en casa?" recibe una probabilidad diferente dependiendo de una estimación experta alternativa. Esto cambia el conocimiento previo, por ejemplo, preguntando al compañero o entrenando a los gatos. Si se diera un caso así, habría que discutir críticamente por qué discrepan los expertos.

Pero quien ahora piense – ¿no debería calcularse siempre una probabilidad de forma *objetiva* y no depender del capricho de las estimaciones de los expertos? – tiene razón, por supuesto, en que los caprichos están fuera de lugar. Pero también tiene que tener claro qué significa objetividad en ciencia. Como se explica en el capítulo 1.1, vivimos en un mundo de verdades relativas, en el que, en términos absolutos, se no

pueden encontrar ni las certezas ni la objetividad con certeza de 100%. Todos los análisis de datos no muestran otra cosa. Más bien existen relaciones causa-efecto increíblemente diversas, algunas de las cuales pueden examinarse más detenidamente con fines científicos y, al mismo tiempo, están aquejadas de un mayor o menor grado de incertidumbre. Las exigencias de objetividad de la estadística clásica socavan en cierto modo esta relatividad cuando van de la mano con una reclamación de absolutez, creando una engañosa sensación de certeza sobre algo que de facto no existe. Esto no significa que no haya soluciones a los problemas que sean mejores que otras soluciones o que incluso puedan ser claramente superiores a ellas. Sólo significa que la ciencia está fundamentalmente cargada de incertidumbres que suele ser mayor ante datos limitados que ante una base de datos completa, sana y robusta y un muestreo excelente. Como debería haber quedado claro en los capítulos anteriores, barreras de significado introducidas convencionalmente o cualquier criterio basado nada más que la convención – y esto no tiene nada que ver con la estadística clásica propiamente dicha, ya que esto se aplica igualmente a la estadística bayesiana más adelante – en el mejor de los casos llevan a una certeza engañosa en la línea de los grupos de presión. Nada de esto tiene que ver con el tema y sus peculiaridades. Sin embargo, la certeza con respecto a la seriedad de las conclusiones debe estar anclada en el tema de la investigación, y ahí es donde los estudios se vuelven muy delgados cuando juzgamos los estudios de esta perspectiva.

El teorema de Bayes no expresa otra cosa, porque tiene en cuenta esta incertidumbre, independientemente de que surja de una base de datos o del juicio de un experto ante la falta de información y la información incompleta. De todas las ciencias, la física, especialmente la astrofísica, la física del plasma y la física cuántica, es la que mejor ha comprendido este principio. La astrofísica trabaja con él porque (Loredo, 1990, 1992) sólo existen unos pocos casos de n = 1 y aún están por investigar. Además, las estrellas no son accesibles a la manipulación experimental, sino sólo a la observación, y esto en relación con periodos de tiempo que superan con mucho a nosotros, los humanos, y al tiempo que presumiblemente nos es dado. La física cuántica ha comprendido esto porque, a más tardar a nivel de átomos y espacios más pequeños, la relación de incertidumbre de Heisenberg (Heisenberg, 1927) ha sustituido completamente la visión newtoniana del mundo: los acontecimientos se definen por probabilidades y no por unidades contables. Por lo tanto, los cálculos estadísticos clásicos sólo son posibles en una medida muy limitada, si es que lo son, para llegar a conclusiones razonables. Más adelante (s. cap. 6.14.6) retomaremos el tema de la subjetividad con el trasfondo de la información cualitativa concreta y las suposiciones sobre la variable "prior" (lo previo).

6.3.2 Caso práctico: otro experimento con té

Una larga charla sobre la cuestión de si a sucesos como un estado de conocimiento previo se les pueden asignar probabilidades sin más cuando todavía no se dispone de datos. Vamos a probarlo. Retomemos el ejemplo anterior del té, esta vez como una modificación fuerte del experimento "té antes de leche o vice versa" según Fisher (1935/1973, véase cap. 4.3.2.1).Queremos averiguar si el té procede de la región de cultivo de Assam bajo la condición de que queramos beber una segunda taza después de una muestra de té y que lo hagamos cuando visitemos cualquier casa de té típica.

6.3.2.1 Problema

Nos interesa la probabilidad posterior $p(Assam \mid 2ª\ taza)$.

.

6.3.2.2 Hipótesis

El cálculo bayesiano de probabilidades requiere que utilicemos un conocimiento previo para el primer cálculo, es decir, la probabilidad a priori de que el té proceda de Assam –

$p(Assam)$.

La Likelihood es la probabilidad de que tomemos una segunda taza de té, si el té procede de Assam, es decir,

$p(2ºAssam \mid Assam)$.

La constante de normalización en el denominador del teorema de Bayes da como resultado la probabilidad (total) de tomar una segunda taza de té, es decir

p(2ª taza).

Las probabilidades correspondientes se derivan de la regla del producto para derivar funciones de la forma f(x) = u(x) * v(x) según f´(x) = u´(x) * v(x) + u(x) * v´(x):

$$p(Assam \mid 2^a\ taza) * p(2^a\ taza) = p(2^a\ taza \mid Assam) * p(Assam) \tag{6.10}$$

y esto lleva al teorema de Bayes con

$$p(\text{Assam} \mid 2^a\ \text{taza}) = \frac{p(2^a\ \text{taza} \mid \text{Assam}) \cdot p(\text{Assam})}{p(2^a\ \text{taza})} \tag{6.11}$$

La diferenciación del denominador de la probabilidad total de una segunda taza de té según la ley de la probabilidad total

$$p(\text{A}) = p(A \cap B) + p(A \cap \bar{B}) \tag{6.12}$$

$$= p(B) \cdot p(A \mid B) + p(\bar{B}) \cdot p(A \mid \bar{B}) \tag{6.13}$$

modifica el denominador:

$$p(\text{Assam} \mid \text{2a taza}) = \frac{p(\text{2a taza} \mid \text{Assam}) \cdot p(\text{Assam})}{p(\text{2a taza} \mid \text{Assam}) \cdot p(\text{Assam}) + p(\text{2a taza} \mid \overline{\text{Assam}}) \cdot p(\overline{\text{Assam}})} \tag{6.14}$$

$$p(\text{Assam} \mid \text{2a taza}) = \frac{p(\text{2a taza} \mid \text{Assam}) \cdot p(\text{Assam})}{p(\text{2a taza} \mid \text{Assam}) \cdot p(\text{Assam}) + p(\text{2a taza} \mid \overline{\text{Assam}}) \cdot p(\overline{\text{Assam}})} \tag{6.15}$$

Debe quedar claro que para la probabilidad global de una segunda taza de té, tanto la condición el té viene de Assam como el té no viene de Assam deben combinarse con el suceso condicional de una segunda taza de té.

6.3.2.3 Conocimientos previos (las Priores)

Pasemos primero al conocimiento previo, que en nuestro caso procederá de la estimación de expertos, ya que no existen datos empíricos de estudios anteriores. Por supuesto, se puede afirmar "el 34% del té procede de China" o "el 90% es Darjeeling" y calcular la diferencia equivalente a uno. Sin embargo, esto no es mucho mejor que adivinar y, sobre todo, sólo demuestra un pensamiento científico muy modestamente desarrollado. En el caso de una falta total de información, a menudo se sugiere en el caso continuo tomar un prior no informativo (Studer, 1996b) para expresar el estado de conocimiento "no tenemos ni idea, porque podría ser de una manera o de otra". Técnicamente, se toman entonces distribuciones uniformes (distribuciones iguales) o distribuciones que no prefieren en modo alguno un rango de valores (por ejemplo, Jeffreys-Carnap, Studer, 1996b o ciertas formas de la distribución beta, véase también la Fig. 6.68, p.735). Nos aventuramos a desarrollar por argumentación una prior informada, aunque no óptima, que entra en la ecuación como conocimiento previo. Una *prior informada* es una distribución o suposición que reduce justificadamente la gama de valores, en principio infinita, y por tanto proporciona una dirección, es decir, crea una ponderación.

El juicio de un experto no se basa en conjeturas, sino, en el mejor de los casos, en una reconstrucción precisa de la información disponible. La estadística de Bayes debe entenderse siempre como un tratamiento de la información (Jaynes, 2003) y toda información tiene en sí misma una distribución de probabilidad, un espacio multidimensional en el que, dependiendo de la situación y el contexto las ondas de probabilidad (pueden) cambiar en función de la situación y el contexto. El punto de vista de la teoría de la información adquirió un estatus significativo sobre todo a través de la obra de E.T. Jaynes (por ejemplo, 1958, 1957a) y

basándose en la teoría de la información del matemático y criptógrafo Claude E. Shannon (1916-2001) sobre la transmisión de señales y la comunicación ocupan un lugar importante dentro de la estadística bayesiana. Por razones pragmáticas, los argumentos que ahora se exponen no son exhaustivos, sino que se limitan a demostrar la forma de pensar y proceder. Nos centramos en determinar la probabilidad (total) de que un té sea originario de Assam aquí en Europa Central o no – *p(té originario de Assam)* – suponiendo que esto puede ayudar a explicar qué té se nos pone delante cuando visitamos (según nuestro ejemplo ficticio) a cualquier casa de té sin pedir un tipo específico de té. Para ello necesitamos algunas condiciones de contorno, de las que sólo enumeramos algunas:

- Se omiten las pequeñas plantaciones de té individuales y las zonas de cultivo específicas porque son demasiado pequeñas y específicas. La atención se centra en las grandes regiones productoras de té. En primer lugar, se distingue entre té negro, verde y oolong (= semifermentado). Una información previa disponible en las casas de té visitadas es que siempre hay sólo té negro. Obtenemos esta información de forma ficticia, porque todas estas casas de té lo anuncian. Así que, simplificando mucho, Japón y Corea se omiten casi por completo, al igual que partes de China, Birmania y una pequeña parte de Darjeeling y Assam, y la única región productora de té de Suiza, porque estas zonas producen casi exclusivamente té verde u oolong y no té negro.
- Nos enteramos de que el té que nos sirven es asequible. Así que se omite el carísimo té europeo de Escocia, ¡una verdadera lástima! El raro e igualmente caro té negro japonés también se cae.
- La cuestión de la época de cosecha (first flush, second flush) es insignificante, ya que no tiene nada que ver con la región de cultivo del té en sentido estricto.
- El té negro bueno y común en Asia procede de China (sabe terroso, a veces ahumado), India (Assam, Darjeeling, a veces del suroeste como Nilgiri, donde hay pequeñas cantidades de té verde, pero sobre todo se produce té negro) y Sri Lanka (éste es originalmente Assam trasplantado a Sri Lanka, es decir, a las tierras altas de UVA). Recientemente se ha producido buen té en África (Kenia), aunque con menos frecuencia en Europa, pero con un alto volumen de exportación. El Nilgiri desempeña un papel subordinado en Europa. Lo mismo ocurre con Indonesia, Vietnam, Sudamérica y Rusia, que exportan té negro en mayor o menor medida, pero desempeñan un papel secundario en el mercado europeo y especialmente el mercado alemán.
- De otro dato, esta vez de un conocedor de todas las casas de té, se deduce que se trata de té puro y no de una mezcla. Así pues, las distintas mezclas de té negro (escocés, inglés, frisón oriental, chai/té de especias, Earl Grey, etc.) desaparecen, donde regiones productoras de té tradicionalmente distantes (por ejemplo, China y Darjeeling, Ceilán y Assam, etc.) se mezclan para crear nuevas variaciones. Éstas suelen dominar los supermercados y también están ampliamente disponibles en tiendas de té especiales.
- Ahora tenemos que evaluar la casa de té más de cerca para ver si hay indicios que apunten a una u otra preferencia por una región de cultivo, más allá de Assam. Una vez más, sabemos por nuestra información privilegiada que en todas las casas de té se sirve Assam, pero no en qué medida. Es decir, Assam es básicamente un té que se puede servir en todas las casas de té.
- No sabemos con más precisión. Si asumimos el principio del azar, es decir, que intervienen tantas influencias diferentes que no podemos acotar de antemano la región de cultivo, entonces a cada región se le asignaría 1/4, es decir, 25%, Assam, Darjeeling, China y Sri Lanka. Pero no es tan sencillo. Un reparto equitativo de las probabilidades parece improbable. Este sería un argumento en contra de una prior ingenua y desinformada. Es posible, pero no utilizaríamos la información y eso, conociéndola mejor gracias a la investigación, sería imprudente.
- En el caso de probabilidades desigualmente distribuidas, habría que investigar cómo se van a evaluar las regiones, ya sea en términos de volumen de exportación o de importancia en el mercado mundial, si abastecen a su propio país o exportan principalmente, etc. La producción mundial y las exportaciones mundiales no dependen de la limitada importación a Alemania. Y esto podría desempeñar un papel importante para las casas de té y qué variedades de té prefieren comprar.
- Una búsqueda de exportaciones a Alemania sólo arroja inicialmente países y no regiones productoras de té en relación con el volumen de exportación (Actualitix, 2016; Deutscher Teeverband, 2017, solo condicionalmente datos visibles: Statista, 2019). Tras cribar las tablas correspondientes, las desiguales probabilidades entre los países parecen muy plausibles. A lo largo de estas fuentes, un orden de países para las importaciones de té a Alemania (DeutscherTeeverband, 2017, 2018). De esta tabla, primero necesitamos solo el valor para la India (2015: 20:91%, 2016: 24:47% y 2017: 27:11%, es decir, x = 24:5%, mediana = 25:47%). Dado que este valor incluye las regiones de Assam, Darjeeling, Dooars y Nilgiri, ahora necesita una estimación de la cuota relativa de estas regiones en el contexto de las importaciones procedentes de la India. Tras nuevas investigaciones (Berliner Teesalon, 2017), nos acercamos a las regiones indias significativas de cultivo de té (por orden alfabético): Assam, Darjeeling, Dooars, Nilgiri, Himal Pradesh, Orissa, Sikkim y Teral. Posteriormente, se

desprende que la región de Assam representa alrededor del 60% de las exportaciones de té de India. Así, esperamos que Assam tenga una probabilidad del 25:47% 60% = 15:28%. Esto significa que, asumiendo la seriedad de las cifras comunicadas, esperamos un té con al menos un 15:28% de probabilidad de la región de Assam. Utilizamos el término "al menos" porque parece plausible que el valor sea superior y la estimación da un límite inferior. La razón es que el té negro en Alemania se asocia generalmente con Assam o Darjeeling y luego domina ligeramente Assam, además de las mezclas que (véase más arriba) no desempeñan aquí ningún otro papel. Una visita a diversas tiendas de té, tiendas de alimentos ecológicos y supermercados muestra que las regiones productoras de té negro se encuentran muy raramente fuera de la India y China, y que en lo que respecta a la India destacan Assam y Darjeeling y muy raramente Nilgiri.

- Nuestra incertidumbre resultante con esta estimación experta se refiere entonces a la cuestión de si el volumen de exportación e importancia de las regiones mencionadas coinciden con la opinión de las casas de té, o si los respectivos gerentes tienen orientaciones diferentes, por ejemplo porque las preferencias locales de sus clientes no se corresponden con las de toda Alemania. Esto podría investigarse en una encuesta, cosa que no haremos.
- Si aparecen nuevos datos inesperados – por ejemplo, sobre la cuota del té de Assam en Alemania o que una casa de té no sólo ofrece té negro, sino también verde u oolong – las probabilidades deben modificarse en consecuencia, según las fuentes mencionadas, que deben considerarse fiables. Tras unas cuantas rondas de pruebas con té y una cantidad creciente de datos, poco a poco podría formularse una hipótesis estable y sólida sobre la procedencia del té de las casas de té. Entonces, el empirismo iría sustituyendo poco a poco a la subjetiva valoración inicial, y así es como debería ser.

Esto es, a grandes rasgos, lo que podría parecer una argumentación y los procesos de pensamiento asociados a ella, que en caso de que falte información intenta formarse un juicio a partir de la información contextual disponible. Por un lado, la argumentación debería mostrar que no es en absoluto trivial llegar a una valoración seria de probabilidades cuando no se dispone de una base de datos. Pero si es así, no queda más remedio que construir una argumentación seria y prudente que tenga en cuenta todas las perspectivas. En este ejemplo ficticio, la cuestión era de dónde proceden los tés de una selección de casas de té. ¿Es ahora "censurable" este planteamiento por ser "subjetivo"? Desde nuestro punto de vista definitivamente: ¡No! La argumentación está bien fundada, utiliza fuentes externas serias e intenta incluir la información disponible. Si falta información o es incompleta, sigue siendo necesario formarse un juicio. ¿Qué alternativas hay a adivinar ciegamente o no hacer nada en absoluto cuando sabemos que existe información fiable? Aparte de intentar relacionar cuidadosamente la información, no hay más alternativas que adivinar. Si se desarrolla una situación de datos empíricos, no hay duda de que hay que utilizarlos como base de datos principal; y así, en la mayoría de los casos, la conjetura inicial es cada vez menos importante e influyente.

Tabla 6.1: *Té (exportaciones en toneladas y porcentaje)*

La comparación con el enfoque de la investigación cualitativa muestra pocas diferencias. Allí también se clasifica la información (por ejemplo, los textos), se relacionan entre sí, se comparan, se comprueban hipótesis, etc., y esto se hace sin la formación de frecuencias (por ejemplo, en el análisis secuencial). E incluso la investigación cualitativa "sufre" de vez en cuando la acusación de subjetividad, aunque esto ha cambiado considerablemente en las ciencias sociales actuales. Los métodos de investigación cualitativa han llegado a la corriente dominante, al igual que los métodos mixtos (Tashakkori & Teddlie, 2003b).

Una discusión comparable al enfoque cualitativo de la estadística bayesiana en un campo de investigación desconocido puede encontrarse en O'Hagan, Buck, Daneshkhah, Eiser, Garthwaite Jenkinson, Oakley & Rakow (2006) o Studer (1996b). Urban M. Studer, físico matemático de la ETH de Zúrich, describe el enfoque bayesiano en la evaluación general de las tasas de recuperación en la terapia de la adicción. Se trata de transferir los conocimientos priores disponibles a una forma numérica en el curso de la evaluación de un centro de terapia de la adicción en régimen de hospitalización (Studer, 1995, 1996a, 1998), aunque en aquel momento no se dispusiera de datos. La razón es que en aquel momento la recogida de datos para la evaluación de la institución, que sólo se había fundado dos años antes. En cuanto al contenido, las consideraciones sobre las perspectivas de éxito en la terapia de la adicción (por ejemplo, ¿es posible el éxito en general? ¿Es posible el fracaso en general?) se combinan de tal forma que se puede esbozar de forma fundamentada el ámbito potencial de la recuperación de la adicción y posteriormente acotarlo.

6.3.2.4 Laplace y la aplicación del teorema de Bayes

En el presente estudio de la casa de té, el empirismo se orienta hacia la condición de la segunda taza de té, que se considera significativa para la pregunta de qué región de cultivo procede un té. Si se trata de una pregunta significativa o no, es una cuestión discutible. No se trata de eso, sino de cómo proceder. Si un estudio es acertado y perspicaz se decide a otro nivel, es decir, cuando se da un paso atrás y se reflexiona sobre todo el proceso. Ahora que se han formulado los conocimientos previos (las priores), se pueden recopilar datos, porque no hay nada más convincente que la experiencia y el empirismo. Para explicarlo mejor, tomemos algunos datos ficticios para el posible resultado de la primera ronda del experimento. Importante es que aquí nos abstenemos de contar, sino que nos limitamos a dar probabilidades de cómo podría ser. En el siguiente ejemplo sobre pruebas médicas se trabaja entonces con frecuencias, por lo que la ecuación de probabilidades con frecuencias relativas requiere una cierta definición de la probabilidad (véase el capítulo 4.2.1), tal como se utiliza especialmente en la estadística frecuentista.

Por lo tanto, damos la vuelta y nos adentramos en la historia de la estadística. En su famoso artículo, Laplace (1814) describe un sistema matemático de razonamiento inductivo con la ayuda de explicaciones probabilísticas. Al hacerlo, deduce el teorema de Bayes sin nombrarlo como tal. Los siete primeros principios que expone son

1. La probabilidad es la relación entre los sucesos ocurridos y todos los posibles, también conocida como fórmula de Laplace:

$$p(A) = \frac{\text{número de sucesos de A}}{\text{número de todos los sucesos posibles}} \qquad (6.16)$$

Este término de probabilidad es frecuentista, es decir, contable, pero puede aplicarse a cálculos bayesianos discretos.

2. Se supone que cada suceso tiene la misma probabilidad. Si no es así, hay que aclarar las probabilidades de cada suceso. La probabilidad total es entonces la suma de las probabilidades individuales de los sucesos que se han producido y de los sucesos que no se han producido.

3. Para los sucesos independientes entre sí, las probabilidades pueden multiplicarse entre sí.

4. Para los sucesos interdependientes A y B, la probabilidad compuesta de los sucesos A y B viene dada por el producto de la probabilidad del suceso A por la del suceso B si se ha producido el suceso A, es decir

$$p(A \bigcap B) = p(A \mid B) \cdot p(B) \qquad (6.17)$$

5. La probabilidad de A dado B es la probabilidad de la ocurrencia de de A y B dividida por la de B, es decir

$$p(A \mid B) = \frac{p(A \bigcap B)}{p(B)} \qquad (6.18)$$

Esto corresponde a una transformación de la regla 4.

6. Aquí Laplace llega al teorema de Bayes con corolarios. A cada una de las causas de un suceso observado se le asigna una probabilidad bajo el supuesto de que el suceso es constante y las probabilidades no cambian. La probabilidad de una de estas causas corresponde a un cociente formado por el numerador con la probabilidad del suceso observado causado por esta causa dividido por la suma de las probabilidades relativas de todas las causas, es decir, la probabilidad total. Si estas causas no son igualmente probables a priori, es necesario sustituir el numerador. Se introduce entonces la probabilidad del suceso multiplicada por la probabilidad de su causa. Esta última corresponde a la probabilidad a priori. Todo ello da lugar al teorema de Bayes con el suceso Ai_e (A_1, A_2, ..., A_n), que especifica las posibles causas del suceso B con p(B) = p(A_1, A_2,...,A_n) y probabilidad a priori p_{prior}(A) para el suceso A:

$$p(A_i \mid B) = \frac{p(B \mid A_i) \cdot p_{\text{prior}}(A)}{\sum_j p(B \mid A_j) \cdot p_{\text{prior}}(A_j)} \qquad (6.19)$$

7. Existen otras derivaciones de estas reglas, por ejemplo, en la regla de sucesión, (rule of succession) donde Laplace trata de la estimación del éxito frente al fracaso bajo el supuesto de que poco o nada se sabe a priori sobre los respectivos resultados. Si *s* es el número de éxitos y *n* el número de intentos, así como n - s el número de fracasos, entonces la fórmula es

$$p(\text{éxito en el siguiente intento}) = \frac{s+1}{n+2} \qquad (6.20)$$

La explicación es que aquí existe el conocimiento previo de que tanto el éxito como el fracaso (pueden) ocurrir realmente. Esto requiere que se haya observado al menos un éxito y al menos un fracaso, de modo que la relación entre el éxito y el número de intentos sea 1/2, que se suma por *s* y *n* en la fórmula anterior. Laplace era consciente de este absurdo y probablemente sólo utilizó el ejemplo porque era fácil de entender. Quizás también estaba aburrido de los ejemplos de la moneda, la urna y la lotería que él mismo utilizaba. Laplace aplicó con gran éxito este pensamiento a una gran variedad de problemas de estadística médica, testimonios ante los tribunales y mediciones astronómicas. Hacia el final de su vida, Laplace trabajó probablemente de forma frecuentista, aunque la distinción entre bayesiano y frecuentista que se utiliza hoy en día ni siquiera debería aplicarse a este periodo. Laplace estaba mucho más preocupado por la aplicación de principios, sin apelar a ningún tipo de idealización o paradigma.

Es importante ver que Bayes se ocupa de probabilidades y no de frecuencias. En el caso discreto, en principio se pueden contar los casos, pero ya no en el caso continuo, e incluso el caso discreto no cambia el hecho de que el concepto bayesiano de probabilidad no es el de la estadística frecuentista. Aquí la definición de la unidad de investigación difiere de los respectivos enfoques. Por supuesto, una p = 0.1 o 10% no es única. Puede ser 10 de cada 100 o 1 de cada 10 o 33 de cada 330. La *tasa base* siempre desempeña un papel decisivo (Gigerenzer, 1991). El concepto bayesiano de probabilidad es, de forma abreviada,

aproximadamente idéntico al lo que utilizamos *intuitivamente* en la vida cotidiana: Estimamos algo y no lo contamos necesariamente, ni las tasas base ni ninguna otra cosa. Pero sí incluimos nuestras experiencias, por ejemplo cómo fue algo ayer o lo que hemos oído, lo que podemos sospechar, inferir, etc. Por supuesto, podemos contar acontecimientos de la experiencia, pero es poco probable que sean representativos. En principio, podemos llegar a una probabilidad para cada suceso – que esto sea cierto es algo completamente distinto.

¿De qué datos disponemos para nuestro ejemplo (`ptII_quan_Bayes_intro-BayesTheorem_tea.r`)?

```
> # prior knowledge
> p.TeeAssam.t0 <- 0.1528
> p.NOT.TeeAssam.t0 <- 1-p.TeeAssam.t0
> # empirical data
> p.2ndcup.cond.TeeAssam.t0 <- 0.7
> p.2ndcup.cond.NOT.TeeAssam.t0 <- 0.55
> # complementary probs
> p.NOT.2ndcup.cond.TeeAssam.t0 <- 1-p.2ndcup.cond.TeeAssam.t0
> p.NOT.2nd.cup.cond.NOT.TeeAssam.t0 <- 1-p.2ndcup.cond.NOT.TeeAssam.t0
> # application Bayes Theorem
> # discrete case
> p.TeeAssam.cond.2ndcup.t0 <- p.2ndcup.cond.TeeAssam.t0 *
                + p.TeeAssam.t0 /
                + (p.2ndcup.cond.TeeAssam.t0 *
                + p.TeeAssam.t0 +
                + p.2ndcup.cond.NOT.TeeAssam.t0 *
                + p.NOT.TeeAssam.t0)
> p.TeeAssam.cond.2ndcup.t0
[1] 0.1866927
```

Para el conocimiento previo, es decir, las suposiciones que se hacen antes de cribar los datos empíricos, se aplica lo siguiente según lo anterior:

- $p(\text{Assam}) = 15.28\,\%$
- $p(\overline{\text{Assam}}) = 100\,\% - 15.28\,\%$

A partir de las visitas a las casas de té, la recogida de datos arroja los siguientes datos empíricos –

- *p(2ª taza | Assam)* = 70%, es decir, normalmente en 7 de cada 10 casos se desea una segunda taza de té si procede de Assam.
- *p(2ª taza |no de Assam)* = 55%, es decir, normalmente en 5 de cada 10 casos se solicita una segunda taza de té, aunque no proceda de Assam.

Debería ser obvio que estas dos probabilidades condicionales no suman a uno. Son probabilidades condicionales cualitativamente diferentes. Por lo tanto, puede ser que siempre queramos una segunda taza de té, tanto si es de Assam como si no. Del mismo modo, puede que sólo queramos una segunda taza en los casos en que el té proceda de Assam, o viceversa, sólo si el té no procede de Assam. Y luego está la zona intermedia, como si tiráramos los dados o dejáramos que nuestro estado de ánimo o el tiempo decidieran.

Un acontecimiento no determina el otro. Ese sería el caso de la probabilidad total para un té de la región de Assam *p(Assam)* y *p(no de Assam)*, que juntas suman uno, ya que se trata de dos acontecimientos complementarios que se excluyen mutuamente y sólo hay dos resultados posibles. Para las probabilidades condicionales anteriores, las probabilidades complementarias son

- p(2ª taza | Assam) = 100% - 70% = 30%
- p(2ª taza | no de Assam) = 100% - 55% = 45%

La aplicación del teorema de Bayes (ptII_quan_Bayes_intro-BayesTheorem_tea.r) conduce a

$$p(\text{Assam} \mid 2^{\text{a}}\ \text{taza}) = \frac{p(2^{\text{a}}\,\text{taza} \mid \text{Assam}) \cdot p(\text{Assam})}{p(2^{\text{a}}\,\text{taza} \mid \text{Assam}) \cdot p(\text{Assam}) \quad + \quad p(2^{\text{a}}\,\text{taza} \mid \overline{\text{Assam}}) \cdot p(\overline{\text{Assam}})} \tag{6.21}$$

$$= \frac{0.70 \cdot 0.1528}{0.70 \cdot 0.1528 + 0.55 \cdot 0.8472}$$

$$= 0.187$$

Se puede interpretar esta probabilidad posterior como así: En caso de que se quiera una segunda taza de té, este té viene con una probabilidad de 18.7 % de Assam.

> Recordatorio 6.1: Bayes – Aprendizaje a partir de la experiencia
>
> Conocimiento previo ahora = conocimiento posterior en el momento anterior de la elicitación

6.3.2.5 Replicación y actualización

El experimento se repite para introducir el aprendizaje a partir de la experiencia (véase más arriba) – que es posible gracias al teorema de Bayes – y para "mostrar una replicación en absoluto". Esta vez, sin embargo, el punto de partida del conocimiento previo es el resultado (probabilidad posterior) del experimento anterior, a saber, p = 0.187. Dado que el conocimiento previo (prior) se define como el conocimiento que existe sobre un tema antes de que se vean los datos empíricos, se puede utilizar el conocimiento posterior del experimento anterior directamente como conocimiento previo del tiempo del experimento actual. Los siguientes valores resultan bajo el supuesto

- p(Assam) = 18.70%
- p(Assam) = 100% - 18.70% = 81.30%

Las nuevas visitas a las casas de té, esta vez con una persona diferente, arrojan nuevos datos empíricos.

- p(2ª taza j Assam) = 90%
- p(2ª taza j Assam) = 35%

A la vista de los nuevos datos, el teorema de Bayes proporciona ahora un estado de conocimiento actualizado a (`ptII_quan_Bayes_intro-Teorema de Bayes_tea.r`)

$$p(\text{Assam} \mid 2^{\text{a}}\ \text{taza}) = \frac{0.90 \cdot 0.187}{0.90 \cdot 0.187 + 0.35 \cdot 0.813} \tag{6.22}$$

$$= 0.372$$

Y esto es lo que hacemos en R (`ptII_quan_Bayes_intro-BayesTheorem_tea.r`):

```
> # replication and update the Bayes Theorem
>
> # new prior knowledge = posterior from previous timepoint
> p.TeeAssam.t1 <- p.TeeAssam.cond.2ndcup.t0
> p.NOT.TeeAssam.t1 <- 1-p.TeeAssam.t1
>
> # new empirical data
> p.2ndcup.cond.TeeAssam.t1 <- 0.9
```

```
> p.2ndcup.cond.NOT.TeeAssam.t1 <- 0.35
>
> p.TeeAssam.cond.2ndcup.t1 <- p.2ndcup.cond.TeeAssam.t1 *
                             + p.TeeAssam.t1 /
                             + (p.2ndcup.cond.TeeAssam.t1 *
                             + p.TeeAssam.t1 +
                             + p.2ndcup.cond.NOT.TeeAssam.t1 *
                             + p.NOT.TeeAssam.t1)
> p.TeeAssam.cond.2ndcup.t1
[1] 0.3711741
>
> # ratio first time point versus t1
> # = ratio posteriors from t0 to t1
> ((.9*.187)/(.9*.187+.35*.813))/((.70*.1528)/(.70*.1528+.55*.8472))
[1] 1.990684
> p.TeeAssam.cond.2ndcup.t1 / p.TeeAssam.cond.2ndcup.t0
[1] 1.988155
```

Obviamente, los gustos y también el origen del té han cambiado. La probabilidad posterior inicial de p = 18.70% resulta ahora p = 37.2%. Pero si nos fijamos en los detalles, en la replicación el favoritismo por una segunda taza de té (en el caso de Assam) ha aumentado del 70% al 90% (Δ = +20%), mientras que el mismo valor (en el caso de no té de Assam) ha bajado del 55% al 35% (Δ = - 20%). Por lo tanto, no es sorprendente que el origen del té de Assam en el caso de una segunda taza de té sea una duplicación de la tasa, a saber, 0.372 / 0.187 = 1.99. Este proceso de actualización, que transforma una posterior en una prior (anterior) para su uso en un nuevo conjunto de datos, podría en principio continuar sin fin. Con muchos más datos, esperamos que en el futuro los valores empíricos por encuesta y considerados individualmente puedan diferir significativamente entre sí (= Likelihood), pero la probabilidad posterior (resumida) cambie cada vez menos hasta alcanzar un valor robusto y persistente con poca fluctuación. Los grandes cambios en la probabilidad posterior son realistas al principio de la recogida de datos, como se ve aquí. A largo plazo, sin embargo, debería producirse una cierta constancia. Cualquier otra cosa no sería aprender de la experiencia. Lo interesante es que, con la estadística bayesiana, la probabilidad posterior final si se analiza todo el conjunto de datos a la vez o sucesivamente mediante el proceso de actualización descrito. Se trata de una implicación del aprendizaje de la experiencia y del razonamiento coherente subyacente (Cox, 1961). Un ejemplo empírico de la tasa de aprobados en el tratamiento hospitalario de la drogadicción demuestra este proceso (véase el capítulo 6.2). proceso (véanse los capítulos 6.8.1.3 y 6.15.2).

En la práctica, la estadística de Bayes es, por supuesto, mucho más complicada y exigente desde el punto de vista matemático. El caso discreto descrito se mantiene muy simple, minimalista y puede calcularse cómodamente a mano. Si tratamos cuestiones más complejas con muchas variables y con el caso continuo, se deben dar distribuciones a priori para cada variable, que marcan el estado actual del conocimiento. Esto no es ni mucho menos trivial, sino exigente y, desde luego, no siempre exacto. Como ya se ha subrayado, el conocimiento a priori debe surgir del objeto. Para determinar el denominador de la probabilidad total, se suelen utilizar complejos algoritmos de simulación MCMC (Gelman, Carlin, Stern & Rubin, 2003, Parte 3), ya que el denominador del teorema de Bayes rara vez puede determinarse analíticamente mediante integración y, por tanto, no existe una solución computacionalmente sencilla. Sólo se hacen excepciones para casos sencillos como el problema Behrens-Fisher de medias iguales y desviaciones estándar iguales (véase el capítulo 6.15.3.8, Bretthorst, 1993) o para la comparación de proporciones (véase el capítulo 6.15.3) o en el caso de distribuciones conjugadas.

Sin embargo, el pensamiento básico no cambia. El teorema de Bayes sigue siendo núcleo fijo y todos los problemas abordados bayesianamente trabajan con él. El procesamiento de la información por el teorema de Bayes sigue siendo coherente y toda la información disponible encuentra su camino en un modelo que siempre da como resultado una probabilidad posterior que determina el grado de verosimilitud del suceso condicional que nos interesa. Para ilustrar esta lógica básica de pensamiento necesitamos

- Conocimiento previo basado en datos anteriores o en argumentos y estimaciones de expertos,
- Likelihood basada en datos empíricos, y
- la probabilidad total basada en todo el espacio de probabilidad.

Ahora repetimos todo con las pruebas médicas anunciadas. Esto demuestra de forma impresionante que las probabilidades son a veces menos intuitivas de lo que uno espera. Con una comprensión cada vez mayor del teorema de Bayes, las razones de esto se hacen más claras y uno llega a conclusiones menos ciegas e irreflexivas en la vida cotidiana sobre los posibles resultados de los acontecimientos.

6.3.3 Caso práctico del diagnóstico médico

Las pruebas en el diagnóstico médico son demostraciones clásicas de Bayes (Stein, 1999-04-23). Y los lectores valientes pueden probarlo de inmediato en la próxima visita al médico.

6.3.3.1 Definición del problema

Si una prueba médica da un resultado positivo, ¿cuál es la probabilidad de estar realmente afectado por la enfermedad analizada? Expresado en probabilidades, es p(enfermedad + | prueba +) o abreviado p(E + | P +).

6.3.3.2 Hipótesis

Las pruebas médicas tienen dos características: Sensibilidad y Especificidad.

Sensibilidad = correcto Positivo = probabilidad de que una prueba indique correctamente un resultado positivo,
por ejemplo, si una persona padece una determinada enfermedad. Una sensibilidad alta es importante cuando se quiere *excluir* una enfermedad con la máxima certeza. Esto se debe a que si se identifica correctamente a todas las personas que padecen la enfermedad, un resultado negativo de la prueba significa que la enfermedad puede excluirse con un alto grado de probabilidad.

Especificidad = correcta Negativa = probabilidad con la que una prueba indica correctamente la ausencia de enfermedad,
por ejemplo, si una persona no padece una determinada enfermedad y está sana en este sentido. Una especificidad elevada es importante si se quiere *detectar* una enfermedad con la máxima certeza. Así, si todas las personas sanas se clasifican correctamente con un resultado negativo, un resultado positivo significa con alta probabilidad la presencia de la enfermedad analizada.

Dado que tanto la sensibilidad como la especificidad no alcanzan exactamente el 100 % en la práctica, existen otras dos probabilidades complementarias relevantes en función del resultado de la prueba:

Falsos Positivos = probabilidad de que alguien no tenga la enfermedad pero la prueba muestre un resultado positivo. Parece que la persona está enferma. Sin embargo, no está enferma, sino sana.

Falsos Negativos = probabilidad de que alguien esté enfermo, pero la prueba no lo muestre correctamente. Parece que la persona está sana. Sin embargo, no está sana, sino enferma.

Los cálculos reales dependen de las tasas base (Krämer, 1998, cap. 7, palabra clave: error de prevalencia, Gigerenzer, 1991), es decir, de la incidencia real (prevalencia) de la enfermedad respectiva que se va a analizar en la población o en un subgrupo, así como del tamaño de la muestra original a la hora de determinar los valores de especificidad y sensibilidad. Las cifras de prevalencia son significativas, ya que una afirmación "el 4% de esta población padece la enfermedad xyz" significa poco si no sabemos si nos referimos al 4% de 4 000 o de 4 000 000 personas. Los enunciados porcentuales sin las tasas de base son valores poco fiables, ya que sólo representan potencias relativas normalizadas: 1 de cada 2 personas corresponde exactamente al 50%, pero también 350 de cada 700, 667 de cada 1 334, etcétera.

Estas cuatro probabilidades corresponden al esquema de cuatro campos que ya hemos aprendido en estadística clásica en el contexto de los tipos de error Tipo I y Tipo II (véase la Tabla 4.2, p.78).

Tabla 6.2: *Diagnósticos médicos (Realidad vs. Resultado de la prueba)*

Realidad		Resultado de la prueba (P)		
Enfermedad		positivo	negativo	Σ
	Sí	$p(E+ \mid P+)$	$p(E+ \mid P-)$	$p(E+)$
	No	$p(E- \mid P+)$	$p(E- \mid P-)$	$p(E-)$
	Σ	$p(P+)$	$p(P-)$	1

Para calcular la probabilidad de enfermedad en el caso de un resultado positivo, examinamos más detenidamente las probabilidades anteriores y utilizamos los conceptos ya introducidos, es decir, las probabilidades incondicional y condicional.

Como ejemplo de una probabilidad incondicional *p*, tomemos la enfermedad y la salud, respectivamente, que son disyuntivamente complementarias de una probabilidad global de uno: si todo el mundo estuviera sano, no habría enfermedades y viceversa:

- Prevalencia = *p(enfermedad)* = incidencia total de una enfermedad en la población. Si la prevalencia es muy baja – es decir, sólo la padecen muy pocas personas – un resultado positivo en una prueba sólo tiene un valor predictivo bajo de que una enfermedad esté realmente presente en el caso de un resultado positivo, porque, de todos modos, sólo unas pocas personas de la población padecen la enfermedad.
- Sano = *p(no enfermedad)* = 1 - *p(enfermedad)*

Por ejemplo, podemos preguntarnos por la probabilidad condicional de padecer una enfermedad si el resultado de la prueba es positivo o por la probabilidad de padecer una enfermedad a pesar de que el resultado de la prueba sea negativo. En general, los elementos de las probabilidades condicionales no son fácilmente intercambiables:

$$p(\text{suceso A} \mid \text{suceso B}) \neq p(\text{suceso B} \mid \text{suceso A}) \qquad (6.23)$$

Para el ejemplo del diagnóstico médico necesitamos cuatro probabilidades condicionales (véase la tabla 6.2).

Sensibilidad = p(enfermedad presente | resultado positivo de la prueba) = $p(E+ \mid P+)$
falsos Positivos = p(enfermedad presente | resultado negativo de la prueba) = $p(E+ \mid P+)$
Especificidad = p(ausencia de enfermedad | resultado negativo) = $p(E- \mid P-)$
falsos Negativos = p(ninguna enfermedad presente | resultado positivo de la prueba) = $p(E- \mid P-)$

Tabla 6.3: *Teorema de Bayes (cálculo abstracto de diagnósticos médicos)*

Realidad		Resultado de la prueba		
Enfermedad		positivo	negativo o falsos negativos	Σ
	Sí	Sensibilidad		Enfermo = Prevalencia = sensibilidad + falsos negativos
	No	falsos positivos	Especifidad	Sano = Población–Prevalencia = falsos positivos + especifidad
	Σ	Sensibilidad + falsos Positivos	falsos Negativos + Especifidad	Población = Sensibilidad + falsos positivos + falsos negativos + especifidad

6.3.3.3 Información empírica

Concretemos la abstracción con números. Tenemos la siguiente información ficticia sobre una enfermedad xyz:

- Σ = En una muestra representativa de la población se someten a la prueba N = 15 000 personas.
- p(E+) = Por término medio, se encuentran enfermas N_1 = 9 personas.
- p(E+ | P+) = La prueba es muy sensible, de modo que por cada 100 personas con la enfermedad, N_2 = 99 de ellas son identificadas correctamente como portadores de la enfermedad ("E+ | P+"; "verdaderos positivos").
- p(E- | P+) = En las personas sanas, la prueba señala una media de N_3 = 3 personas falsamente como portadoras de la enfermedad ("falsos positivos").

Los cálculos de la tabla 6.4 se aplican según la tabla 6.2.

6.3.3.4 Aplicación del teorema de Bayes

Si, según el ya conocido teorema de Bayes, p(A | B) = p(B | A) p(A) = p(B) y la probabilidad total en el denominador p(B) = p(B | A) p(A) + p(B | A-) p(A-) se cumple: p(A) = prevalencia, p(B + | A+) = sensibilidad y p(B- | A-) = especificidad, se puede aplicar el teorema de Bayes para calcular p(A | B). Para simplificar, podemos escribir una pequeña función R. Dadas son la prevalencia, la sensibilidad y la especificidad (`ptII_quan_Bayes_intro-Teorema de Bayes_diagnósticomédico.r`).

```
# persons tested
sampleN <- 15000
illness.pos <- 9
sensitivity <- 99/100
false.positive <- 3/100
```

Ahora tenemos solamente entrar los valores y mostrar los resultados:

Tabla 6.4: *Teorema de Bayes (ejemplo numérico de diagnóstico médico)*

	Abrev.	Datos (N)	Cálculos	Ejemplos
Población	N	15 000		
Enfermos		9		
Sensibilidad (Prueba)	$p(E+ \mid P+)$	99/100		
falsos Positivos	$p(E- \mid P+)$	3/100		
Prevalencia			Enfermos/Población	9/15000= 6e-04
Sanos			1–Prevalencia	1–6e-04= 0.9994
falsos Negativos	$p(E- \mid P+)$		1–Sensibilidad	1–0.99= 0.01
Especifidad	$p(E- \mid P-)$		1–falsos Positivos	1–0.03= 0.97

```
# calculations
prevalence <- illness.pos/sampleN
healthy <- 1-prevalence
false.negative <- 1-sensitivity
specifity <- 1-false.positive
```

Entonces sabemos los valores de salida:

```
> prevalence
[1] 6e-04
> healthy
```

```
[1] 0.9994
> false.negative
[1] 0.01
> specifity
[1] 0.97
```

La probabilidad posterior p(A|B) de padecer realmente la enfermedad si el resultado de la prueba es positivo es del 1.94%. Es decir, aproximadamente 2 de cada 100, porque

```
> # 0.99*6e-04 / (0.99*6e-04 + (1-0.97)*(1-6e-04))
> BayesTheorem(p.A=9/15000, p.BcondA=0.99, p.NOTBcondNOTA=0.97)
p(A|B) p(B)   p(B|A) p(A)   p(!A)  p(B|!A) p(!B|!A)
0.0194 0.0306 0.9900 0.0006 0.9994 0.0300  0.9700
> BayesTheorem(p.A=illness.pos/sampleN,
                  + p.BcondA=sensitivity,
                  + p.NOTBcondNOTA=1-false.positive)
p(A|B) p(B)   p(B|A) p(A)   p(!A)  p(B|!A) p(!B|!A)
0.0194 0.0306 0.9900 0.0006 0.9994 0.0300  0.9700
> BT.res1.medtest <- BayesTheorem(p.A=prevalence,
                  + p.BcondA=sensitivity,
                  + p.NOTBcondNOTA=specifity)
> BT.res1.medtest
p(A|B) p(B)   p(B|A) p(A)   p(!A)  p(B|!A) p(!B|!A)
0.0194 0.0306 0.9900 0.0006 0.9994 0.0300  0.9700
```

Como variante, reducimos la tasa base en un factor de 10 y, por tanto, aumentamos masivamente la prevalencia, es decir, la aparición total de la enfermedad en la población. Todo lo demás permanece constante. Es de esperar que la probabilidad posterior aumente ahora significativamente porque la incidencia de la enfermedad es mayor con la misma calidad de las pruebas. Un cambio en la tasa de base necesitaría en realidad un cambio en la especificidad y la sensibilidad, que son propiedades de la prueba en sentido estricto. Las probabilidades resultantes para p(E + | P+) dependen directamente de esto.

```
> # lower the base rate by factor 10
> fac <- 10
> BayesTheorem(p.A=9/15000*fac, p.BcondA=0.99, p.NOTBcondNOTA=0.97)
p(A|B) p(B)  p(B|A) p(A)  p(!A) p(B|!A) p(!B|!A)
0.166  0.036 0.990  0.006 0.994 0.030   0.970
> BT.res2.medtest <- BayesTheorem(p.A=prevalence*fac,
                  + p.BcondA=sensitivity,
                  + p.NOTBcondNOTA=specifity)
> BT.res2.medtest
p(A|B) p(B)  p(B|A) p(A)  p(!A) p(B|!A) p(!B|!A)
0.166  0.036 0.990  0.006 0.994 0.030   0.970
> diff.fac <- BT.res2.medtest["p(A|B)"]/BT.res1.medtest["p(A|B)"]
> names(diff.fac) <- "fac.diff"
> diff.fac
fac.diff
8.6
> # =
> .1661074/.019427
[1] 8.6
> # 8.550337
```

Reducir la tasa de base en un factor de 10 conduce a un aumento de la probabilidad posterior de 0.1661074 / 0.019427 = 8:55, es decir, ¡la probabilidad posterior de tener la enfermedad si la prueba da un resultado positivo se ha multiplicado por más de ocho! Y esto sin que haya pasado nada con la calidad de las pruebas. Es "sólo" la prevalencia la que ha aumentado debido al cambio en la tasa de base. Esto debería demostrar claramente la dependencia de las probabilidades de sus tasas de base. La calidad de las pruebas – ya sea para el caso positivo o negativo – depende directamente de la tasa de base y no existe independientemente de ella.

Por supuesto, el cálculo puede realizarse en una notación diferente. Cada cual puede decidir por sí mismo si la formulación abstracta con las letras a y b o una designación mejor le parece más propicia para su propia comprensión.

- PR = Prevalencia = p(A) / 9 = 15000
- SE = Sensibilidad = p(B + | A+) = 0.99
- SP = Especificidad = p(B + | A-) = 0.97
- EPT = enfermedad en caso de resultado positivo = SE * PR / [SE * PR + (1 - SP) * (1 - PR)]

```
> 0,99*6e-04 / (0,99*6e-04 + (1-0,97)*(1-6e-04))
[1] 0.019
> # persons ill (=correct) if diagnosis = positive
> 0.019427*15000
[1] 291.405
```

En relación con la muestra actual, N = 0.019427 * 15000 = 291.405, es decir, normalmente 292 personas se clasifican correctamente como portadoras de la enfermedad si dan positivo, es decir, dan positivo y realmente tienen la enfermedad. Esta lógica no es inmediatamente intuitiva, ya que no nos limitamos a sobreestimar heurísticamente las probabilidades según la estructura de este ejemplo, sino que debemos vincular de forma lógica y coherente distintas cantidades entre sí, ya que las cantidades respectivas no resultan simplemente de forma directa de las demás o no pueden sumarse en uno. Según el teorema de Bayes, se aplica lo siguiente:

$$p(A|B) + p(B|A) \neq 1 \tag{6.24}$$

De forma equivalente, podemos imaginar cuestiones de pedagogía y psicología (por ejemplo, sobre pruebas y diagnósticos) o de la judicatura (por ejemplo, sobre la cuestión de si los acusados son culpables a la vista de pruebas más o menos convincentes sobre el trasfondo de las probabilidades condicionales). Esto es exactamente lo que se ha hecho y se hace a menudo, sobre todo en combinación con los cuatro campos para integrar diferentes perspectivas (Beck-Bornholdt & Dubben, 2003, cap. 14; Krämer, 1998, p.165.; Dubben & Beck-Bornholdt, 2005, cap. 5). En el ejemplo anterior, hemos examinado probabilidades discretas con dos variantes disyuntivas cada una, por ejemplo, enfermedad y salud, pruebas positivas y negativas, etc. En principio, esta lógica puede complicarse a voluntad, tanto para las discretas tanto para el caso discreto como para el continuo. En este último caso, entran en juego distribuciones de probabilidad continuas y ya no sumas de probabilidades discretas para calcular las probabilidades totales. La cosa se complica aún más cuando el modelo contiene varias variables, algunas de las cuales interactúan entre sí. Esto da lugar a una estimación compleja del modelo estadístico.

Por tanto, no sólo podemos estimar los resultados positivos y negativos de las pruebas. De hecho, las posibilidades son ilimitadas. Por ejemplo, podemos investigar el rango de residencia de los cuantos, comparar las tasas de éxito en terapia o hacer predicciones sobre las trayectorias educativas y sus factores de influencia. Según la tradición, Laplace investigó si una moneda es "justa" o cuál es la masa del planeta Saturno (Gunther, 2013). En este último caso, Laplace sólo se equivocó en un 0.367%. Así, Laplace da la masa de Saturno como fracción de la masa del Sol y llega a un valor de 3512 (Pulskamp, 2019). Las mediciones oficiales de la NASA son 3499,1, ¡una proeza para la época en torno a 1815! Investigaciones similares fueron realizadas por Jaynes (1963) sobre la equidad de un cubo (van Enk, 2014-08-28), sólo que mucho más extensas en las afirmaciones derivables (Studer, 1996b, véase Cap. 6.14.5), basadas en el Principio de Máxima Entropía (ver Cap. 6.14).

Estas cuestiones tienen en común que se basan en el teorema de Bayes como mínimo común denominador. Por este motivo, todas estas realizaciones de preguntas de investigación son representantes de la estadística bayesiana, es decir, cuando el resultado es una probabilidad posterior está disponible para su discusión.

En otro ejemplo numérico (Greuel, s.f.) de diagnóstico del cáncer, a la prevalencia se le asigna un valor de 6 = 1000, a la sensibilidad 0.98 y a la especificidad 0.96.

```
# cancer example
# data given
prevalence <- 6/1000
sensitivity <- .98
specifity <- .96
```

Las derivaciones para los sanos, los falsos positivos y los falsos negativos resultan coherentes con

```
# calculations R-Code
healthy <- 1-prevalence
false.positive <- 1-sensitivity
false.negative <- 1-specifity
```

y conducen a

```
> healthy
[1] 0.994
> false.negative
[1] 0.04
> false.positive
[1] 0.02
```

Tarea 6.1: Prueba del SIDA

La prueba Elisa (Enzyme Linked Immuno Sorbent Assay) es un procedimiento de cribado y examina el SIDA. A su vez, existen los dos indicadores de sensibilidad, que indica el SIDA si una persona está infectada.

(`ptII_quan_Bayes_intro-BayesTheorem_medicaldiagnosis.r`).

Esta es del 99,9%. El segundo indicador es la especificidad, que no da un resultado positivo si una persona no está infectada. En el caso de la prueba Elisa, es del 99,5%. Estas cifras parecen indicar un alto grado de certeza en la predicción. Sin embargo, sabemos que esto depende de la prevalencia, es decir, de la frecuencia de la infección por SIDA en un determinado grupo de población. Actualmente, 84 700 personas en Alemania (a finales de 2016) tienen SIDA (Wikipedia, 2019a). En Alemania viven 81. 390.400 personas (countrymeters, 2019). Qué probabilidad hay de tener SIDA si la prueba es positiva, es decir, p(E + | P+)?

Cambia los valores de prevalencia o tasa base y de sensibilidad y especificidad. ¿Corresponde el resultado del teorema de Bayes a las expectativas subjetivas debidas de los cambios introducidos?

La aplicación del teorema de Bayes a esta base de datos de Greuel da ahora como resultado

```
> BT.res.cancer <- BayesTheorem(p.A=prevalence,
                  + p.BcondA=sensitivity,
                  + p.NOTBcondNOTA=specifity)
> BT.res.cancer
p(A|B)    p(B)      p(B|A)    p(A)      p(!A)     p(B|!A)   p(!B|!A)
0.1288344 0.0456400 0.9800000 0.0060000 0.9940000 0.0400000 0.9600000
> BT.res.cancer["p(A|B)"] * prevalence
p(A|B)
0.0007730061
> 1-BT.res.cancer["p(A|B)"]
p(A|B)
0.8711656
```

Ahora la pregunta sería, ¿cuál es más probable? ¿O tener realmente cáncer

```
# in case of positive test, what is more probable?
# cancer is real
cancer.real <- sensitivity * prevalence
```

o no tener cáncer?

```
> # no cancer, cancer is not real
> false.negative * (1-prevalence)
[1] 0.03976
> # =
> cancer.notreal <- false.negative * healthy
> # = MAP hypothesis = maximum a posteriori
> cancer.real
[1] 0.00588
> cancer.notreal
[1] 0.03976
```

Esto se denomina la hipótesis MAP (Greuel, s.f. diapositiva 8), es decir, la *Máxima Probabilidad A Posteriori* (McElreath, 2015). MAP es un estimador bayesiano especial que estima un parámetro de interés desconocido por el valor modal, es decir, por el máximo de la distribución posterior. Es algo comparable al método de máxima Likelihood (MLE), pero basado en un conjunto de datos diferente. El MLE se basa en la selección del parámetro de forma que, bajo su distribución, los datos observados sean lo más plausibles posible, y este valor del parámetro representa entonces el resultado. Estos valores son iguales si en el cálculo bayesiano se asume una distribución uniforme a priori no informativa, lo que nunca debería ser el caso en la práctica. Sin embargo, la estimación MAP desempeña un papel mucho menor en la estadística bayesiana que, por ejemplo, el estimador ML en la estadística frecuentista, aunque ahora hay libros de texto enteros sobre Bayes y R (McElreath, 2015) que argumentan casi exclusivamente con MAP. Además, con Bayes toda la distribución posterior desempeña un papel y no solo una estimación puntual. Si se utilizan funciones de pérdida en la teoría bayesiana de la decisión, el estimador MAP no es necesariamente el estimador óptimo (Berger, 1985, Sección 4.4). Esto varía dependiendo de la función de pérdida (lineal/absoluta, cuadrática, ...) y de si una ponderación de la función de pérdida es eficaz.

¿Ahora, cómo de grande es la diferencia entre estas dos probabilidades de cáncer sí/no?

```
> # compare both terms
> cancer.real > cancer.notreal
[1] FALSE
> cancer.real / cancer.notreal
[1] 0.1478873
```

Las respectivas probabilidades individuales pueden relacionarse con la suma de las probabilidades de cáncer sí/no

```
> # normalising
> cancer.real / (cancer.real+cancer.notreal)
[1] 0.1288344
> cancer.notreal / (cancer.real+cancer.notreal)
[1] 0.8711656
> # =
> 1-cancer.real / (cancer.real+cancer.notreal)
[1] 0.8711656
```

que, en el caso de la pregunta sobre la probabilidad real de tener cáncer en caso de un resultado positivo de la prueba, corresponde al resultado del teorema de Bayes.

```
> # compare with results of Bayes-Theorem
> BT.res.cancer["p(A|B)"] == cancer.real / (cancer.real+cancer.notreal)
p(A|B)
TRUE
```

A veces, sin embargo, no es importante saber que no se tenía una enfermedad, sino precisamente que ya se tenía. COVID-19 es una de ellas.

6.3.4 Caso práctico: Fiabilidad de una prueba COVID-19

En el contexto de la pandemia de COVID-19 a partir de 2019, en varios países se planteó la cuestión de si debía imponerse a la población una tarjeta de inmunidad general. Esta debería documentar si una persona ya ha sido infectada por COVID-19, independientemente de la presencia de síntomas o de una evolución más o menos grave de la enfermedad. La idea subyacente es que el cuerpo acumula anticuerpos a través del contacto con los virus y puede reactivarlos rápida y específicamente durante la siguiente infección para evitar enfermar a través de la propia respuesta inmunitaria del organismo. Aparte del muy problemático debate sociopolítico de poner "etiquetas de salud" a las personas y posiblemente negarles el acceso a determinados ámbitos de la vida en función de ello, un procedimiento de este tipo presupone una serie de cosas ciertas:

- Una prueba puede identificar con gran precisión los anticuerpos contra el COVID-19 y distinguirlos de otros anticuerpos que se hayan creado contra la gripe u otras variantes del COVID, por ejemplo.
- Por tanto, estadísticamente se espera que los falsos positivos y los falsos negativos sean idealmente cero y que las personas con anticuerpos contra COVID-19 puedan distinguirse claramente de las personas sin anticuerpos contra COVID-19. En la práctica, esto es imposible.
- La detección de anticuerpos COVID-19 significa que se ha adquirido inmunidad a la enfermedad durante el mayor tiempo posible. Esta suposición dista mucho de haberse demostrado empíricamente en la investigación sobre el COVID-19.

Por consiguiente, como siempre, se trata de afirmaciones estadísticas y no de certezas. Los términos especificidad y sensibilidad ya se han introducido en los ejemplos anteriores. Sin embargo, mientras que en el diagnóstico del cáncer, por ejemplo, el objetivo era siempre evitar tener una enfermedad en la medida de lo posible, la situación se invierte con la prueba de anticuerpos COVID-19. En este caso es ventajoso haber tenido una enfermedad para ser lo más inmune posible después. En términos de calidad de una prueba, esto significa:

- Si una persona da falsamente positivo pero no lo es, se crea la ilusión de inmunidad. En consecuencia, el comportamiento de esta persona puede cambiar y, después de todo, contraerá la enfermedad, por ejemplo, a través de un comportamiento descuidado mientras el estado pandémico sigue vigente, u otros.
- Si una persona se somete falsamente a la prueba como negativa, pero no lo es, se crea la ilusión de una falta de inmunidad (dejaremos de lado en este punto la discusión de si la inmunidad puede inferirse de los anticuerpos en absoluto, ya que esto no tiene nada que ver con la prueba de anticuerpos). Una consecuencia sería que esta persona podría ser menos capaz de participar en determinados aspectos de la vida (por ejemplo, profesionalmente), lo que puede acarrear diversos problemas (aislamiento social, reducción de las oportunidades profesionales, desempleo, etc.).
- Esto demuestra que tanto los falsos positivos como los falsos negativos tienen consecuencias directas en la vida cotidiana.

Ahora bien, lo que nos interesa es la probabilidad de haber estado enfermo si el resultado de la prueba es positivo. Esta probabilidad puede derivarse del teorema de Bayes, pero requiere necesariamente información sobre la distribución de COVID-19 en la población. Los criterios de calidad de la prueba por sí solos no son suficientes. Se puede imaginar así:

- Si una enfermedad no se da en absoluto en la población, es decir, nadie está infectado, la probabilidad de tener esta enfermedad a pesar de un resultado positivo en la prueba es muy baja. Esto se debe a que la probabilidad global es simplemente muy baja independientemente de las pruebas y es cero en el caso límite.
- Sin embargo, si la prevalencia de una enfermedad en la población es muy alta, porque todo el mundo está infectado, incluso con un resultado negativo en la prueba, la probabilidad de tener la enfermedad es bastante alta. La razón es que la probabilidad global en la población es tan alta y en el caso límite es uno.

- Como puede verse, la probabilidad de tener una enfermedad depende no sólo del resultado de la prueba, sino a un nivel más amplio de la probabilidad de propagación de esta enfermedad en la población. La prueba sólo proporciona una probabilidad condicional.

Pasemos al ejemplo de los datos. Tomemos los datos sobre la precisión de la prueba de anticuerpos COVID-19 de la empresa CeGaT GmbH, con sede en Tubinga. En su página web (a fecha 02.07.2020), indican el isotipo de anticuerpos IgG de la prueba, así como una sensibilidad del 94.4% y una especificidad del 99.6%. Además, se indica que en N = 23 casos reactividad cruzada con otros coronavirus patógenos humanos no pudo observarse.

En primer lugar, construimos una función R basada en la sensibilidad, la especificidad y la probabilidad de distribución p en la población. Aquí nos guiamos por Schneider, Dinant y Szecsenyi (2006) y adoptamos – esperemos que tras una prueba exhaustiva (`ptII_quan_Bayes_intro-BayesTheorem_covid-19-test.r`) – la fórmula impresa (ibid., p.122) expresada en R:

```
# function to calculate posterior prob
posterior.test <- function(sensi, speci, p)
{
sensi*p / (sensi*p + (1-speci)*(1-p))
}#
data
# characteristics of the test
# specifity
speci1 <- 0.996
# sensitivity
sensi1 <- 0.944
```

La propagación en la población es dinámica. Por lo tanto, en el contexto de la pandemia, podemos suponer que las cifras crecerán. Dependiendo de las restricciones de contacto con el público, las cifras crecen más o menos rápidamente. Además, hay un cierto número de casos no notificados que son desconocidos por naturaleza y a veces es muy difícil de estimar indirectamente. Las proyecciones basadas en los puntos calientes de COVID-19, como Ischgl en la provincia de Tirol (Austria), Kreis Heinsberg en el distrito administrativo de Colonia (Alemania), etc., pueden considerarse como un límite superior de un número no notificado en el supuesto de que los respectivos instrumentos de análisis sean fiables y no añadan más incertidumbre.

El 19 de mayo de 2020, la prevalencia de COVID-19 en Alemania era de 180 000 personas, según la OMS. Es decir, alrededor del 0.225% o p = 0.00225. Pongamos esto en la fórmula

```
# overall probability in population R-Code
p1 <- 0.0025
post1 <- posterior.test(sensi=sensi1, speci=speci1, p=p1)
post1
# =
0.944*.0025/(.944*.0025+(1-.996)*(1-.0025))
# 0.3716535
# =
BayesTheorem(p.A=.0025, p.BcondA=0.944, p.NOTBcondNOTA=0.996)
```

El resultado es 37.165%. Esto significa que si la prueba es positiva, hay un 37% de probabilidades de haber padecido realmente la enfermedad, siempre que la tecnología de medición subyacente funcione. Ahora nos interesa la curva de cómo cambia la probabilidad a medida que aumenta la prevalencia en la población.

```
sek <- seq(0,1, length.out=100+1) R-Code
probs <- posterior.test(sensi=sensi1, speci=speci1, p=sek)
probs
```

Se puede mostrar eso de forma gráfica (véase Fig. 6.1):

```
plot(sek*100,probs*100, bty="n", col="darkred",
```

```
        type="l", pre.plot=grid(),
        main="COVID-19 in case of positive test result",
        xlab="Distribution in population (%)",
        ylab="COVID-19 p(%)")
```

Si ahora nos exigimos al menos un 90% de probabilidad subjetiva de haber padecido la enfermedad, entonces esto es posible – suponiendo una especificidad y sensibilidad constantes – si x% de la población ya ha tenido COVID-19. Este valor x se puede calcular:

```
> # percent population required to acquire
> # a certainty of x% in the test
> perc.sec <- c(0.5, 0.9, 0.99)
> N.total <- 83200000
> for(i in perc.sec)
+ {
+   perc.prop <- sek[probs >= i][1]*100
+   cat("\n% security (COVID-19 test) = ",i,
+          "\n% in population required (base rate): = ",
+          perc.prop,"\n",sep="")
+ cat("people required = ",perc.prop/100*N.total,"\n",sep="")
+ abline(v=perc.prop, col="blue", lty=2)
+ text(perc.prop+1, i*100, pos=4,
+          labels=paste(perc.prop,"%",sep=""),
+          cex=2)
+ }
% security (COVID-19 test) = 0.5
% in population required (base rate): = 1
people required = 832000
% security (COVID-19 test) = 0.9
% in population required (base rate): = 4
people required = 3328000
% security (COVID-19 test) = 0.99
% in population required (base rate): = 30
people required = 24960000
```

En ese caso, al menos el 4% de la población debe haber padecido ya COVID-19. Con 83.200.000 personas (2020), esto significa 3.200.000 personas. Si uno es paranoico y necesita subjetivamente al menos un 99% de probabilidad, necesitamos en consecuencia al menos una prevalencia del 30% en la población o alrededor de 25 millones de personas. Este cálculo no distingue según el sexo, la edad, el grupo de riesgo, etc., sino que es un cálculo general no específico.

Ahora podríamos repetir el cálculo según Bayes, porque para estar seguros tendríamos que someternos dos veces a la prueba. Pero esto supone que los errores de la prueba no están correlacionados. Si hacemos la misma prueba dos veces, seguramente no será así, porque los errores de la prueba – si es que los hay – son de naturaleza sistemática. Por ejemplo, podría ocurrir que una proteína de la sangre, que también se encuentre en el virus, pero en este caso no provenga del virus sino de otra fuente (cuál de ellas no es importante aquí). Esto sería un falso positivo, cuya probabilidad es constante y, por tanto, se da igual en la primera que en la segunda prueba. Mejor sería una prueba diferente con una tecnología diferente.

Para mayor claridad, tomamos no obstante la misma prueba y volvemos a poner en la misma fórmula la probabilidad posterior de la primera medición como prior:

```
> # replicate with initial post1 value
> post2 <- posterior.test(sensi=sensi1, speci=speci1, p=post1)
> post2
[1] 0.9928871
```

Alternativamente, se podría reescribir la fórmula de la siguiente manera

```
> posterior.test2 <- function(sensi2, speci2, p2)
+ {
```

```
+ sensi2^2*p2 / (sensi2^2*p2 + (1-speci2)^2*(1-p2))
+ }
> # =
> posterior.test2(sensi2=sensi1, speci2=speci1, p2=p1)
[1] 0.993
> # =
> BayesTheorem(p.A=post1, p.BcondA=0.944, p.NOTBcondNOTA=0.996)
p(A|B) p(B)  p(B|A) p(A)  p(!A) p(B|!A) p(!B|!A)
0.993  0.353 0.944  0.372 0.628 0.004   0.996
```

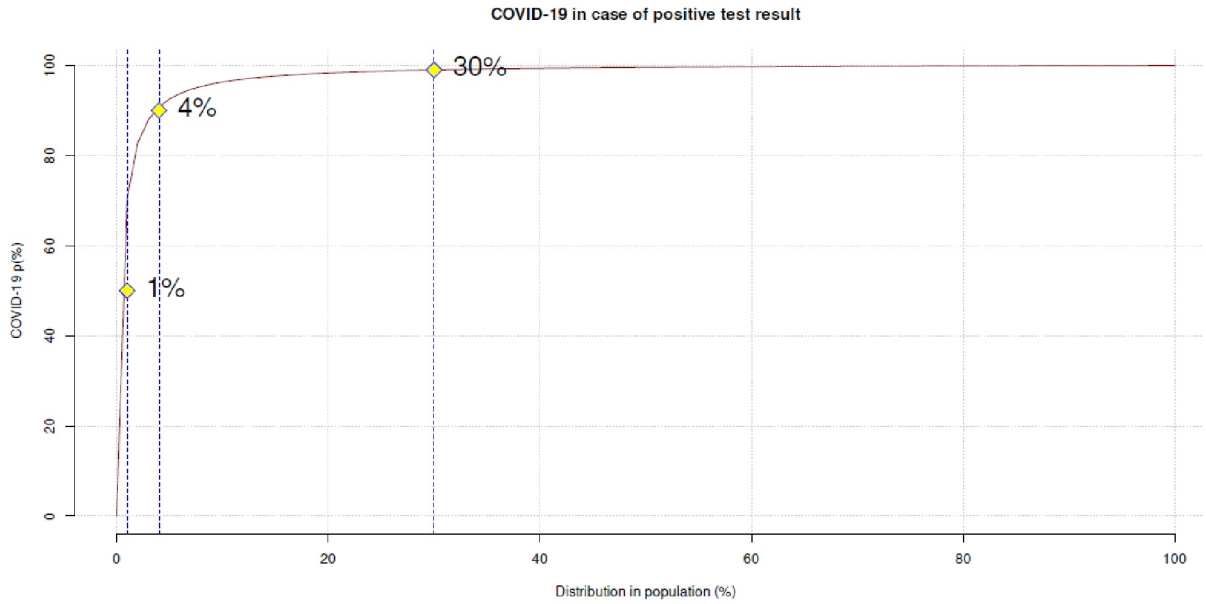


Figura 6.1: *Fiabilidad de una prueba COVID-19 (distribución requerida en la población).*

Esto es posible porque entre dos pruebas idénticas sucesivas no cambian ni la sensibilidad ni la sensibilidad y el valor *p* para la distribución en la población puede manejarse como constante. En sentido estricto, por supuesto, *p* cambia con una cierta probabilidad después de una prueba positiva, pero dado el tamaño de la población esto es insignificante aquí.

Si supusiéramos que nuestra prueba no contiene errores estructurales, llegaríamos a un nivel del 99.21% con dos pruebas y, por tanto, al mismo nivel que si el 30% de la población ya estuviera infectada (véase más arriba). En la práctica, podemos orientarnos del siguiente modo si una prueba de anticuerpos resulta relevante con las restricciones antes mencionadas (por ejemplo, los anticuerpos son en realidad una expresión de inmunidad:

- Buscamos dos pruebas que tengan unos buenos valores de especificidad y sensibilidad aproximadamente comparables, pero que preferiblemente se basen en una tecnología diferente (por ejemplo, identificación de proteínas),
- hacemos ambas pruebas y
- comparamos ambas pruebas (el orden no importaría) con la prevalencia en la población en el momento de la prueba según el teorema de Bayes.

Con un poco de suerte, puedes arreglártelas con dos pruebas para obtener una alta probabilidad de un resultado razonable. Los costes de la prueba de la empresa CeGaT se estiman en 25.- (a 02.07.2020); por lo tanto, 2 pruebas cuestan el doble y otras pruebas probablemente estarán en una categoría de precio similar.

Tarea 6.2: Cantidades y sus relaciones

La tarea para los lectores consistiría en realizar estos cálculos para diferentes países, tasas de crecimiento y pruebas (es decir, con diferentes valores de sensibilidad y especificidad) para hacerse una idea de la relación entre estas cantidades. Salga de inmediato de las curvas y marque los intervalos de probabilidad de certeza o incertidumbre subjetiva que sean relevantes para usted.

Sin embargo, no se debe reducir la estadística de Bayes al caso discreto. En general, la realidad rara vez es discreta, pero las curvas de probabilidad pueden encontrarse en todas partes, por lo que el caso continuo modela mejor la realidad.

Tras varios ejemplos de aplicación del caso discreto, ahora daremos un paso atrás para examinar más de cerca el propio teorema de Bayes. Existen varias interpretaciones para entenderlo mejor.

6.4 El teorema de Bayes

6.4.1 Fondo histórico

Mientras que la estadística clásica cuenta y trata la probabilidad de los datos a la vista de una hipótesis nula formulada de forma más o menos específica o en el contexto de la teoría de Neyman-Pearson (entonces con hipótesis alternativa), la estadística de Bayes se ocupa de varias probabilidades en paralelo, incluida la ya conocida de la estadística clásica: la Likelihood o conocida allí como $p(datos \mid H_0)$. El principio de Likelihood establece que los datos de una muestra generada por la recogida de datos y la información que contienen están completamente representados por la función de Likelihood $L(\theta|x)$. Sin embargo, el teorema de Bayes va más allá, ya que permite inferir de *p(datos | hipótesis)* a *p(hipótesis | datos)*. Para ello, siguen siendo necesarias las probabilidades incondicionales *p(datos)* y *p(hipótesis)*. La unión de estas cuatro probabilidades da lugar al *teorema de Bayes*, que debe su nombre al sacerdote Thomas Bayes (1701-1761), cuya principal obra, "Essay Towards Solving a Problem in the Doctrine of Chances" sólo se publicó póstumamente en 1763 (Bayes, 1763). Stigler (2013) ofrece detalles sobre la publicación. El documento no contiene el teorema de Bayes como ecuación matemática, y el único retrato del reverendo Thomas Bayes en Internet (un dibujo) probablemente ni siquiera sea del propio Bayes (TheIMSBulletin, 1988). Así pues, hay muchas cosas que no están claras en lo que respecta al reverendo Thomas Bayes como persona histórica. Su biografía, que puede reconstruirse, puede encontrarse en Bellhouse (2004). El teorema de Bayes fue posteriormente "redescubierto" de forma independiente y publicado por Pierre Simon Laplace (1749-1827); y hay autores que, en vista de la extensa obra de Laplace, lo consideran el verdadero fundador de esta forma de estadística (lukeprog, 2011). En su obra, Laplace describe de forma matemáticamente moderna el sistema de ecuaciones conocido como teorema de Bayes.

Históricamente, es importante que *p(A|B) vs. p(B |A)* o *p(hipótesis | datos) vs. p(datos | hipótesis)* se han denominado a menudo *inversas* entre sí. La "paradójica" relación inversa de estas cantidades entre sí fue discutida por Fisher (1922) junto con el concepto ya existente de *probabilidad inversa*. La probabilidad inversa se refiere a la distribución de probabilidad de una variable no observada. Fisher intentó determinar la probabilidad inversa mediante la inferencia fiducial sin tener que recurrir a una distribución a priori. Aldrich (2000) describe la posición y el uso del concepto de probabilidad inversa en los primeros trabajos de Fisher. El punto de partida de Fisher era que uno no hace afirmaciones de probabilidad sobre parámetros y acabó con el método de elección – máxima Likelihood. Este sigue siendo su concepto central hasta el final de su vida.

Fienberg (2006) va más allá y traza cómo el concepto de probabilidad inversa se convirtió en la probabilidad bayesiana de elección. Así que, históricamente, tenemos diferentes componentes, todos más o menos redondeados en el teorema de Bayes:

- *Estadística inferencial* = determinación de una variable no observada
- *Probabilidad directa* = distribución de probabilidad de una variable observada (este término se fusionó más tarde en la función de Likelihood, que da una *goodness of fit* del modelo estadístico para inferir de los datos observados a los parámetros no observados del modelo basándose en la distribución conjunta de todos los datos observados).
- *Probabilidad prior* = asignación de una distribución de probabilidad esperada a un suceso o variable no observado (por ejemplo, basándose en el conocimiento de expertos, subjetivamente por estimación o basándose en resultados empíricos previos e investigaciones anteriores, que también incluyen probabilidades posteriores más antiguas).
- *Probabilidad bayesiana* ≈ *Método de la probabilidad inversa* = asignación de una distribución de probabilidad a una variable no observada.
- *Likelihood* = distribución de los datos en vista de la variable no observada (no es una distribución de probabilidad, ya que puede tomar valores superiores a uno).
- *Probabilidad posterior* = distribución de una variable no observada cuando se dan tanto los datos (Likelihood) como una distribución de probabilidad a priori. Si se normaliza el producto de probabilidad Prior * Likelihood, se obtiene el resultado del teorema de Bayes. En caso contrario, el numerador de Prior * Likelihood es proporcional a la probabilidad posterior.

6.4.2 Derivación

El teorema de Bayes puede derivarse de (al menos) tres formas diferentes:

- Regla del producto
- Teoría de conjuntos
- Árbol de decisión

Se aconseja a los lectores que elijan la variante que entiendan intuitivamente. Para nosotros la derivación de la regla del producto (véase más arriba) es fácil de entender.

6.4.2.1 Regla del producto

La derivación de la regla del producto de la teoría de probabilidad se describe en detalle en Jaynes (2003). Dos sucesos se ponen en relación entre sí en un nivel probabilístico, de modo que las probabilidades condicionales recíprocas multiplicadas por las respectivas probabilidades individuales incondicionales (totales) dan el mismo resultado.

$$p(A|B) \cdot (B) = p(B|A) \cdot p(A) \tag{6.25}$$

Como se puede ver, no estamos hablando de Prior, Likelihood, Total o Posterior. Se trata de dos sucesos A y B con sus probabilidades individuales incondicionales *p(A)* y *p(B)*. Las probabilidades condicionales *p(A|B)* y *p(B|A)* establecen una conexión directa entre los sucesos, sin introducir en modo alguno el concepto de causalidad, que no desempeña ningún papel en el dominio abstracto de la probabilidad.

Si ahora interesa una de las probabilidades condicionales *p(A|B)* o *p(B|A)* y, de nuevo, no hay preferencia por cuál es la que interesa, basta con reordenar la ecuación de la siguiente manera

$$p(A|B) = \frac{p(B|A \cdot p(A)}{p(B)} \tag{6.26}$$

También se podría haber reordenado la ecuación a *p(B|A)*. Lo que representan A y B se elige arbitrariamente y según el contexto. El denominador *p(B)* resulta como la suma del espacio de posibilidades del numerador *p(B|A) * p(A)*. Para el caso discreto simple, esto es según la ley de la probabilidad total

$$p(B) = [p(B \mid A) \cdot p(A)] + \left[p(B \mid \overline{A}) \cdot p(\overline{A})\right] \tag{6.27}$$

y significa que una normalización debe referirse a todas las manifestaciones de este producto. Para muchos sucesos disyuntivos esto resulta en

$$p(B) = \sum_{j=1}^{N} p(B \mid A_j) \cdot p(A_j) \tag{6.28}$$

Para el caso continuo el sumatorio se convierte en una integral

$$= \int p(B \mid A) \cdot p(A)\, \mathrm{d}A \tag{6.29}$$

$$= \int p(B,\ A)\, \mathrm{d}A \tag{6.30}$$

El teorema de Bayes para el caso continuo viene dado por

$$p(A|B) = \frac{p(B|A) \cdot p(A)}{\int p(B|A) \cdot p(A)\, \mathrm{d}A} \tag{6.31}$$

Ahora el teorema puede llenarse de contenido. El resto es más o menos complejo y complicado. La integral en el denominador del caso continuo no puede resolverse analíticamente para problemas complejos. Por eso se calcula numéricamente con simulaciones MCMC (véase cap. 6.13). Los términos que aparecen en cada caso pueden clasificarse según la Tabla 6.5.

6.4.2.2 Teoría de conjuntos

Desde el punto de vista de la teoría de conjuntos (véase la Fig. 6.2), el teorema de Bayes se deriva de la superposición entre los sucesos A y B y sus probabilidades. No parte de la regla del producto, sino de la probabilidad condicional *p(A|B)*:

$$p(A|B) = \frac{p(A \cap B)}{p(B)} \tag{6.32}$$

Esto significa que la probabilidad condicional *p(A|B)* es el cociente de la intersección de A y B dividido por la probabilidad total de B, es decir, *p(B)*. El numerador de la ecuación puede expandirse mediante la fracción *p(A)/p(A)* a

$$p(A \mid B) = \frac{p(A \cap B \cdot \dfrac{p(A)}{p(A)}}{p(B)} \tag{6.33}$$

$$= \frac{\dfrac{p(A \cap B}{p(A)} \cdot p(A)}{p(B)} \tag{6.34}$$

Ahora aplicamos la ecuación con las letras inversas:

$$p(B \mid A) = \frac{p(A \cap B)}{p(A)} \tag{6.35}$$

Y por el efecto de la simetría

$$p(A \cap B) = p(B \cap A) \tag{6.36}$$

ponemos ahora *p(A | B)* y recibimos la teorema de Bayes conocido:

$$p(A \mid B) = \frac{p(B \mid A) \cdot p(A)}{p(B)} \tag{6.37}$$

Tabla 6.5: *Teorema de Bayes (contenido y significado)*

Abreviatura	Denominación	Explicación
	Estadística inferencial	Determinación de una variable no observada (independiente del método)
p(A)	Probabilidad PRIOR	Asignación de una distribución de probabilidad a una variable no observada (p.ej., sobre la base de conocimientos previos o de estudios empíricos previos)
p(B | A)	Likelihood	Distribución de probabilidad de los datos dada una variable no observada (no se trata de una distribución de probabilidad en el sentido empírico; en la estadística clásica, se maximiza esta función de Likelihood y se utiliza para determinar el valor *p*)
p(B)	Probabilidad total /evidencia	espacio de sucesos de los datos combinados con la probabilidad a priori (en el caso continuo, esto suele resolverse mediante simulaciones MCMC, ya que no puede resolverse analíticamente; en el caso discreto, es la suma de las probabilidades individuales).
p(A | B)	Probabilidad POSTERIOR	Distribución de una variable no observada dados los datos y la distribución a priori

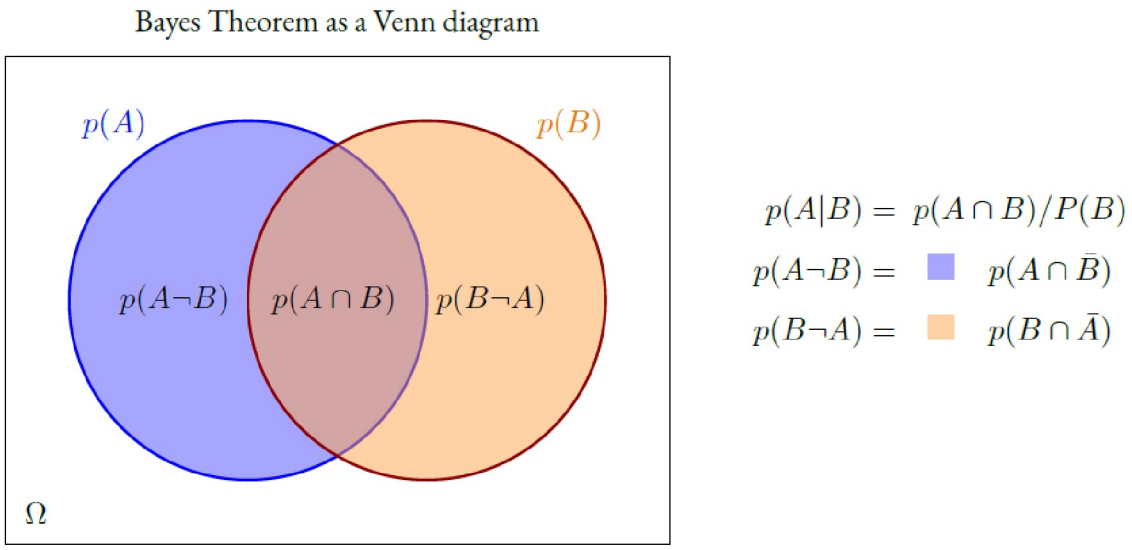


Figura 6.2: *El teoréma de Bayes – Teoría de conjuntos*

6.4.2.3 Árbol de decisión

Los árboles de decisión son adecuados para problemas como las pruebas médicas comentadas, pero resultan menos intuitivos para la variante abstracta del teorema de Bayes. La Figura 6.3 muestra un árbol de sucesos de este tipo con los números del ejemplo de la prueba del cáncer.

Sin embargo, también funciona en abstracto, como muestra la Figura 6.4 (de Wikipedia, 2019b). Como puede verse, se enumeran las intersecciones respectivas de los sucesos A y B, así como las probabilidades individuales resultantes por camino, que son de naturaleza incondicional o condicional y contienen la expresión positiva o negativa del suceso respectivo.

Estas posibilidades están enlazadas combinatoriamente y ordenadas de forma lógica. El conjunto parece bastante abstracto – y lo es. Para un caso con más de dos variables, un diagrama de árbol de este tipo ya sería completamente confuso y difícilmente razonable de interpretar.

Aquí, ya hay probabilidades para 2 variables con 2 valores cada una (caso positivo, negativo) y dos veces un conjunto completo de las probabilidades condicionales a lo largo de los 2 valores mencionados. Esto hace $2+2+(2*2)+(2*2)=12$ probabilidades, de las cuales 4 son incondicionales y $2*4$ condicionales.

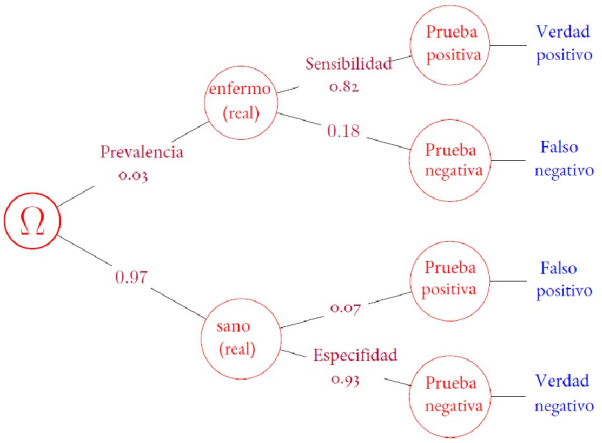


Figura 6.3: *Teorema de Bayes (Àrbol de decisión – prueba del cáncer)*

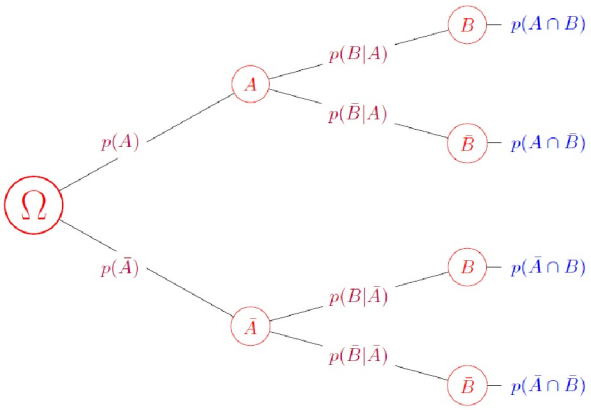


Figura 6.4: *Teorema de Bayes – Árbol de decisión*

6.4.2.4 Diagrama

La figura 6.5 muestra el teorema de Bayes en forma de diagrama, de modo que las conexiones de las distintas perspectivas se ponen de manifiesto en general en la variante simple. En el centro está el espacio de sucesos Ω y a partir de ahí las probabilidades respectivas de que se produzcan los sucesos A y B o de que no se produzcan los sucesos no-A y no-B. Los puntos de esquina contienen entonces las intersecciones de las probabilidades asociadas. Las flechas tienen dos caras para subrayar que el teorema de Bayes como ecuación permite un alto grado de flexibilidad en su aplicación.

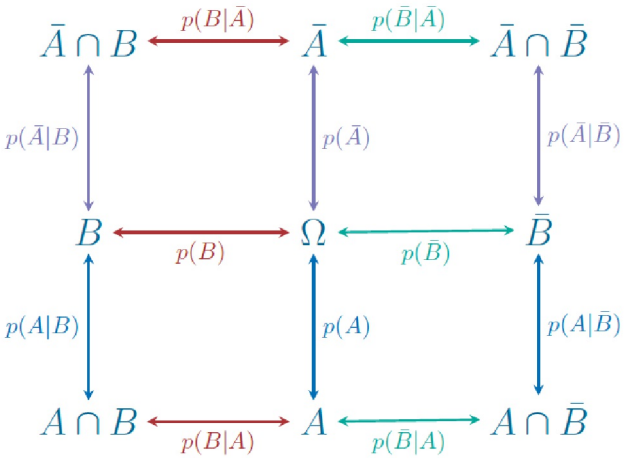


Figura 6.5: *Teorema de Bayes – Diagrama*

6.4.3 Importancia del teorema de Bayes

Los cálculos posibles de este modo contrastan con los de la estadística frecuentista. En el siglo XX, la controversia sobre la cuestión de la relación inversa de $p(A|B)$ y $p(B|A)$ dio lugar a décadas de debate más o menos acalorado, impulsado sobre todo por Fisher y Neyman-Pearson. Una reseña histórica del uso del estadístico de Bayes puede encontrarse en un post en LessWrong a continuación (lukeprog, 2011) y en Jaynes (2003). Por ejemplo, parece ser que Alan Turing utilizó Bayes para desarrollar el hackeo de la encriptación alemana Enigma en la Segunda Guerra Mundial y la Marina estadounidense lo utilizó para intentar encontrar una bomba H perdida o para localizar submarinos rusos. En el transcurso del siglo XX, la estadística de Bayes experimentó un lento renacimiento en la ciencia, empezando por la física (especialmente la física cuántica, la mecánica estadística, la física del plasma, todo ello resumido en McGrayne, 2011) y ahora está cada vez más presente en las ciencias sociales. Es difícil imaginar la estadística sin la estadística de Bayes.

Una de las razones de este renacimiento es que la estadística de Bayes ni siquiera conoce muchos de los problemas de la estadística clásica (por ejemplo, las pruebas múltiples del mismo tipo, la limitación a la probabilidad de los datos, la incapacidad de aprender de la experiencia o de actualizar el estado del conocimiento ante nueva información o datos, etc.), lo que ofrece un enfoque mucho más intuitivo de las probabilidades a la hora de interpretarlas (por ejemplo, los intervalos de confianza bayesianos, véase el capítulo 6.8.4.1). Además de los trabajos de Thomas Bayes y Simon Laplace, los fundamentos lógicos proceden en particular de Pólya (1954a, 1954b), Cox (1961), Jaynes (1957a, 2003) y muchos otros. Según Jaynes (1958, p.1f.), se trata básicamente de aplicar el sentido común en forma de cálculos lógicos y razonados a problemas reales y finitos: "El principio cualitativo de Laplace es su famosa observación de que la teoría de la probabilidad no es más que sentido común reducido a cálculo'. [...] no es sólo un juego de palabras, sino una afirmación literal de un hecho". Para llevarlo a cabo, según Jaynes (2003, cap. 1.7), se necesitan tres desideratas, basadas en el trabajo de Cox (1946):

1. los números reales representan el grado de plausibilidad en forma de probabilidades
2. el sentido común guía el procedimiento
3. coherencia de las conclusiones

La plausibilidad se operativiza mediante la probabilidad o viceversa, la probabilidad es una función de la plausibilidad. Desde el punto de vista de la teoría de la información, el teorema de Bayes expresa el estado actual de conocimiento o error y para un escalado, es decir, una normalización al intervalo de 0 a 1, en el que se definen las probabilidades, se calcula el denominador del teorema de Bayes.

> Recordatorio 6.2: Estado actual de los conocimientos
>
> El estado actual del conocimiento corresponde a un compromiso entre el conocimiento previo y la influencia de los datos empíricos, que denominamos *estado actual del error*.

En general, cuanto mayor es el conjunto de datos empíricos, menor es la influencia del conocimiento previo y los datos determinan el aspecto de la distribución posterior. Sin embargo, también hay casos en los que esto no ocurre (Freedman, 1963, 1965). Según el teorema de *Bernstein-von Mises* (Doob, 1949, véase para una discusión más detallada de las condiciones, Locker, 2009), la distribución de probabilidad posterior es asintóticamente independiente de la anterior a medida que aumenta el tamaño de la muestra y la cantidad de información obtenida de ella. El teorema afirma que en los modelos paramétricos la distribución posterior se distribuye asintóticamente en torno al parámetro verdadero (consistencia del teorema de Bayes). Sin embargo, como señalan Diaconis y Freedman (1986) en su publicación sobre la consistencia de la estimación bayesiana, esta convergencia sólo se aplica condicionalmente. Si la variable aleatoria se encuentra en un espacio de probabilidad infinitamente contable, la variable prior desempeña un papel de apoyo y lo hace independientemente del tamaño de la muestra. La independencia de la posterior y la anterior (prior) ya no se da. Los autores advierten enérgicamente contra una aplicación mecánica del método. La misma advertencia es expresada por Gigerenzer y Marewski (2015), y nuestros comentarios anteriores ya han anticipado esta advertencia. Una prior no es ni una conveniencia matemática ni una necesidad aceptable, sino que requiere una suposición justificada por argumentos y anclada en el objeto de estudio, independientemente de que se disponga o no de datos empíricos. Es concebible, por ejemplo, que se disponga de datos, pero que no se ajusten exactamente a la pregunta. Utilizarlos entonces podría dar lugar a sesgos y, en este sentido, podría darse el caso de confiar más en una o varias estimaciones de expertos que en los datos existentes para formar una hipótesis a priori.

Recientemente, en el contexto de los factores de Bayes (véase el capítulo 6.8.1.4) se ha observado una tendencia a negar prácticamente el conocimiento previo y a seleccionar en función de criterios matemáticos. El artículo de Lortie-Forgues e Inglis (2019) sirve de ejemplo. Aquí, los autores realizan una evaluación a gran escala con muestras grandes y calculan los tamaños del efecto así como los factores de Bayes. La relevancia de estos es basado en la clasificación de Jeffrey (1939/1961). Ni siquiera se discute el hecho de que para un enfoque bayesiano – factores de Bayes o completamente bayesiano con probabilidades a posteriori -- es necesaria una prior bien fundada. Así pues, el enfoque practicado parece descontextualizado y da la impresión de que no es importante omitir componentes de procedimiento importantes a la hora de elegir un instrumento de análisis de datos. Al final del artículo, ante la avalancha de datos, nos preguntamos sobre qué tema han escrito realmente los autores. Además, los factores de Bayes se utilizan principalmente para practicar pruebas de hipótesis nulas y no para modelizar relaciones complejas con una base teórica. No obstante, los autores también proporcionan otros factores de Bayes para hipótesis alternativas adicionales en el sitio web de la revista incluyendo los datos y código R, aunque sin discutirlos con más detalle en lo que respecta a sus implicaciones teóricas y prácticas significativas. El artículo al que hacen referencia los autores (Dienes, Coulton & Heather, 2018) para justificar el cálculo específico de los factores de Bayes se limita a afirmar con respecto a las priores "Tomamos los intervalos de confidencia del 95 % como aproximaciones

de los intervalos de credibilidad correspondientes con priores vagos" (ibíd., p.8). Aparte de eso, no se asigna ningún otro significado a la prior y el término no aparece en absoluto en el artículo. Este enfoque libre de contexto contradice directamente el enfoque bayesiano y la pretensión que Jaynes (2003, p.62, véase también p.196) tiene de un análisis bayesiano de datos,

> „The point is that the best inferences we can make about any phenomenon – whether in physics, biology, economics, or any other field – must take into account all the relevant information we have, regardless of whether that information refers to times earlier or later than the phenomenon itself; this ought to be considered a platitude, not a paradox."

Sin embargo, el conocimiento previo (la variable prior) *siempre* desempeña un papel importante en la estadística bayesiana y a menudo le ha valido la crítica de la *subjetividad*. Sin embargo, desde nuestro punto de vista, no se trata de si el conocimiento previo entra o no en la ecuación, sino de si está justificado, es repetible y reproducible, está orientado al contexto y a la pregunta, se basa en el conocimiento empírico o refleja las estimaciones de los expertos, etc. Especialmente en situaciones de máxima incertidumbre, es importante utilizar toda la información razonable de forma coherente (Studer, 1996b). Resolver una situación así sólo según principios matemáticos parece tener poco sentido.

6.5 Excurso sobre la subjetividad

El teorema de Bayes resulta especialmente valioso cuando se dispone de pocos datos y es necesario transformar el conocimiento cualitativo experto o específico del dominio en una distribución a priori (O'Hagan, Buck, Daneshkhah, Eiser, Garthwaite, Jenkinson, Oakley & Rakow (2006). Esto se suele criticar y rechazar desde una perspectiva frecuentista por ser *subjetivo*, pero debería conllevar un debate sobre lo que significa *subjetivo frente a objetivo* en ciencia (Sprenger, 2018), ya que – véase. Capítulo 6.5.2.4 sobre Bayes o 4.3.7 sobre estadística clásica – faltan criterios absolutos o indubitables de origen natural.

6.5.1 Excurso sobre la física: no está muy lejos lo objetivo de lo subjetivo

Cuando mayor es la pretensión de objetividad y la afirmación asociada de la generalizabilidad de los resultados, más parece que los investigadores implicados a veces "olvidan" o "ignoran" la relatividad de los criterios de verdad científica y en su lugar se pierden en su propia subjetividad e intencionalidad sin ser conscientes de ello; y al mismo tiempo reclaman la máxima objetividad para la validez de su propia posición. Un impresionante estudio de caso lo proporciona un artículo y los posts y entradas de blog asociados de Hossenfelder (2012[a], 2016[a], 2019[a]), un físico teórico que pretende demostrar, de forma puramente teórica y completamente libre de empirismo, que no puede haber libre albedrío humano. Por empíricamente libre, queremos decir que utiliza el bien estudiado modelo estándar de la física, pero no ha realizado ni un solo experimento con un ser humano. Su audaz conclusión salta directamente del modelo estándar de la física a los seres vivos o la humanidad y generaliza completamente a todos los seres humanos muertos, vivos y futuros. Al fin y al cabo, se trata de una hipótesis que puede fallar en principio. Es cuestionable que pueda falsarse en la práctica.

En nuestro excursus, sin embargo, no estamos interesados en si el libre albedrío existe o no. Adoptamos aquí una posición neutral. Sólo nos interesa saber cómo lo justifica el autor. Cuando a continuación se habla de "física", se trata de una mera abreviatura pragmática de las tesis defendidas por el autor y de la pretensión universal asociada al poder explicativo de la física para prácticamente todo. No se refiere al tema de la física con todos los demás físicos. Dado que gran parte de lo que sigue suena como si defendiéramos vehementemente una posición de libre albedrío, debemos recordar que esto se debe a que contradecimos metódicamente al autor. No sabemos cómo es en realidad.

La discusión sobre la cuestión del libre albedrío es, sin duda, social y políticamente relevante, ya que tiene implicaciones significativas, por ejemplo, en los procesos penales, cuando se trata de la cuestión de la culpabilidad y sus factores de influencia, o cuando alguien acude a psicoterapia y quiere cambiar.

Para recapitular, nos remitimos primero a las explicaciones del capítulo [parte:Conocimiento y sabiduría] sobre la teoría de la ciencia - la verdad es relativa: para cada posición existe al menos una contraposición legítima, que por supuesto puede ser inverosímil, pero que en el fondo merece la pena investigar empíricamente. Empecemos con una cita del blog del autor (Hossenfelder, 2016[a]):

> „According to our best present understanding of the fundamental laws of nature, everything that happens in our universe is due to only four different forces: gravity, electromagnetism, and the strong and weak nuclear force. These forces have been extremelywell studied, and they don't leave any room for free will."

A continuación, el libre albedrío se define en el artículo de la siguiente manera (Hossenfelder, 2012[a], p.2):

> „1′) An agent in possession of free will is able to perform an action that does not inevitably follow from all in principle available information at any time preceding the action."

Simplificado, esto significa que con libre albedrío tiene lugar una acción que no puede predecirse por la suma de los estados temporalmente precedentes y la información contenida en ellos. Desgraciadamente, el autor no profundiza en cómo se pueden utilizar los estados físicos cuánticos para sacar conclusiones directas y convincentes sobre lo que ocurrirá exactamente en el mundo y si la cuestión del libre albedrío no está necesariamente ligada a la cuestión de qué es la vida. Porque a partir del punto en el que hay algo más aparte de la materia, la afirmación de reducción anterior deja de ser válida. Puesto que aquí se trata de omnisciencia, es decir, de determinismo, sería interesante ver cómo se justifica esto desde el punto de vista de la teoría científica. Al fin y al cabo, hasta la fecha no existe ningún criterio absoluto de verdad. Hossenfelder (2016[a]) señala „that we don't know how consciousness really works (true but irrelevant). ." y por eso la conciencia es obviamente irrelevante, es decir, un subproducto físico. El autor no demuestra que la conciencia pueda reducirse completamente a la física. No basta con insistir en que sólo existen las cuatro fuerzas de la naturaleza porque han sido muy bien estudiadas. Esto se corresponde aproximadamente con el argumento de mirar sólo en lugares iluminados porque allí hay mucha luz -- y al mismo tiempo suponer que nada puede existir allí donde hay oscuridad, porque en otro lugar hay mucha luz y aquí no.

La base del argumento es, por tanto, que la física está por encima de todas las demás ciencias y puede utilizar un criterio absoluto de verdad, aunque "Según nuestro leal saber y entender [...]" se utiliza como descripción lingüística que restringe de nuevo esta tesis. Sin embargo, junto con la cita anterior "Estas fuerzas han sido extremadamente bien estudiadas [...]", implica que la propia autora no está segura de que todo sea ya conocido, pero actúa como si lo fuera. Partiendo de tal formulación, aún queda espacio suficiente para algo nuevo, aunque estadísticamente tenga una probabilidad previa baja. Sin embargo, la propia autora no se da cuenta de ello. Sólo cita teorías físicas que no pueden explicar necesariamente la vida y las trata de forma determinista. La física ni siquiera puede medir la vida o la conciencia con mayor precisión, y mucho menos definirla. Hossenfelder también ignora que podría haber otras dimensiones físicas, como postuló el físico Bukhard Heim (1905-2001), o campos de información, como defiende Rupert Sheldrake (1942-). Ambos planteamientos tendrían que analizarse empíricamente con detenimiento. Sólo porque a algunos les parezcan absurdos, no hay razón para ignorar el empirismo.

Al intentar aplicar una visión física supuestamente objetiva y empírica a un posible principio universal y generalizarlo así de forma casi atemporal, el autor comete una serie de errores en la teoría de la ciencia. Sólo mencionaremos los más importantes:

- La ciencia trabaja con un criterio de verdad relativo

- Falta de búsqueda explícita de contrapruebas
- Los modelos no son la realidad como tal
- Incrustación en un sistema
- Conclusiones de generalización inadmisibles
- Suposiciones implícitas sin base empírica de datos.

6.5.1.1 Criterio de verdad relativa

En el capítulo 1.1 (El mundo es relativo) analizamos en detalle que la ciencia utiliza un criterio de verdad relativo y no absoluto. Las afirmaciones absolutas repecto a hechos empíricos están fuera de lugar. Por supuesto que se puede hacerlo, pero entonces hay que demostrar que realmente se ha investigado todo el ámbito de posibilidades de tal manera que ni siquiera las investigaciones futuras puedan encontrar nada aquí. La autora ignora este punto porque sólo se refiere a muy pocas teorías y conceptos físicos como el teorema de Bell, la física cuántica como tal, la interpretación de Copenhague de la física cuántica o el demonio de Laplace.

6.5.1.2 Falta de búsqueda de contrapruebas

Según Popper, primero deberíamos intentar falsarnos a nosotros mismos y no reducir nuestro trabajo a la búsqueda de pruebas positivas de nuestras tesis. El autor se refiere exclusivamente a la física y sostiene que no existe ningún modelo explicativo fuera de la física que pueda asociarse al libre albedrío. Al mismo tiempo, no es lo bastante específico como para excluir ámbitos como la vida, la conciencia, etc. por motivos empíricos. Desde nuestro punto de vista, todo se reduce a la cuestión de si la física puede explicarlo todo: la materia viva y no viva.

Tomemos un ejemplo audaz de la espiritualidad. Aquí se relatan con regularidad experiencias en los niveles altos que aparentemente ya no son de naturaleza material o que incluso permiten una experiencia sin tiempo, lo que no niega el hecho de que el cuerpo siga envejeciendo aquí y ahora en el tiempo. Como se explica en el capítulo 1.3 (Los conocimientos dependen de la cultura), el criterio más elevado de conocimiento en las tradiciones espirituales asiáticas prácticas es la experiencia personal directa y no la creencia en instrumentos de medición o razonamientos deductivos. En sentido estricto, para invalidar la validez de un modelo espiritual tan antiguo, la autora tendría que demostrar que tiene tanta conciencia de sí misma que ha experimentado los niveles superiores de espiritualidad con las experiencias asociadas y, al mismo tiempo, puede explicarlos basándose en su propia experiencia, es decir, no cognitivamente. Entonces tendría que llegar a la conclusión de que todas estas experiencias pueden remontarse a causas puramente materiales. Incluso sería legítimo que afirmara: "He alcanzado tales y tales niveles, otros en el mismo nivel pueden confirmarlo, y he descubierto...". La autora no aporta estas pruebas, y puesto que opera en un nivel polivalente, en realidad esperamos que las aporte. Al fin y al cabo, se trata de profundas intuiciones sobre la naturaleza del mundo en relación con nuestro mundo.

Retomemos la experiencia de la atemporalidad, un experimento mental: si alguien experimentara un estado mental de atemporalidad, toda la información del pasado, el presente y el futuro se experimentaría casi en paralelo. Esto significaría que las teorías de Hossenfelder, que se basan en secuencias temporales estrictamente unidireccionales, serían al menos falsificadas por la experiencia, ya que las fuerzas del cambio ya no pueden aplicarse a la conciencia en un estado de atemporalidad. No está claro si entonces existe el libre albedrío, ya que obviamente no queda nada que esté directamente conectado con nuestro mundo. Sin embargo, la atemporalidad implica virtualmente el infinito, lo que no equivale a decir que este otro estado no pueda existir. Las tradiciones espirituales, como la enseñanza de Buda sobre los niveles más elevados, es decir, la experiencia más allá del espacio y del tiempo, no afirman otra cosa.

Es fácil pensar en otros ejemplos que parezcan menos descabellados o absurdos; por otra parte, quien hace afirmaciones omniscientes debe aportar también las pruebas correspondientes de que ha comprobado todo el ámbito de posibilidades y, desde nuestro punto de vista, no basta con hacer afirmaciones cognitivas,

sino que hay que añadir la experiencia personal. De lo contrario, nos encontramos en el nivel de los sistemas de creencias, que es lo que, en definitiva, es la ciencia: creencia en los instrumentos de medida y en su validez, creencia en los resultados de las investigaciones de otros investigadores, creencia en el propio intelecto y en las capacidades cognitivas, etc. Nada de esto se experimenta directamente, sino de forma mediada. Nada de esto se experimenta directamente, sino que se interpreta y se cree a través de los propios órganos sensoriales. Sin embargo, esto no sólo se aplica a la física, sino a todas las ciencias. Esto no es un problema en sí mismo siempre que trabajemos con un criterio de verdad relativa. Sin embargo, desde nuestro punto de vista, sí surge un problema con los enunciados universales. Este punto nos lleva directamente a la función de modelización de la ciencia.

6.5.1.3 Los modelos no son la realidad

Los modelos describen el mundo, pero son "falsos" per se porque no son el mundo en sí, es decir, el objeto de la investigación, sino que sólo proporcionan el lenguaje de la descripción y la explicación y, por tanto, "sólo" tienen un carácter simbólico (véase el modelo semiótico de C.S. Peirce). Por supuesto, los modelos pueden ser bastante pragmáticos, útiles y provechosos, y el empirismo se dirige hacia ellos, pero "el mapa no es el territorio" (Korzybski, 1933[a], 1941[a], véase también Bateson, 1985[a].

La confusión o mezcla de modelo y realidad tiene como consecuencia el desplazamiento de las pretensiones de verdad. La realidad es verdadera como tal, pero nuestra cognición es defectuosa y, por tanto, sólo relativamente verdadera. Históricamente, podemos ver que todas las épocas creían comprender el mundo, pero las épocas posteriores vieron y ven el mundo con otros ojos. Cabe suponer que este cambio continuará y que las generaciones futuras podrán dedicarse a ciencias completamente distintas y llegar a realizaciones diferentes, es decir, investigar cosas que ahora no parecen posibles y que más adelante serán normales. Piense en viajar a la Edad Media con un smartphone y la batería completamente cargada. Entonces probablemente te quemarían por brujo, en otras culturas podrías ser una diosa...

En resumen, no debemos tomarnos demasiado en serio nuestros descubrimientos; cambian.

6.5.1.4 Integración en un sistema

Como explican Whitehead y Russell (1910[a]) en Principia Mathematica sobre la teoría lógica de los tipos, las ambigüedades en el lenguaje pueden llevar a totalidades ilegítimas, lo que puede conducir a un círculo vicioso y crear una paradoja. El resultado son las falacias. Se trata de enunciados que se aplican a sí mismos, o dicho de otro modo: ¿puede un elemento del sistema proporcionar información completa y sin ambigüedades sobre sí mismo mientras sea un elemento del sistema entre muchos otros? Del mismo modo, ¿puede un sujeto hacer afirmaciones universales sobre sí mismo y sobre todos los demás sujetos, es decir, sobre el mundo en su conjunto, basándose únicamente en su propia comprensión de la física como profesión? ¿Sería entonces tal afirmación una expresión del libre albedrío o no, y puede analizarse esto en absoluto si no existe? ¿Es entonces la decisión de afirmar la inexistencia del libre albedrío una decisión "libre" o simplemente una consecuencia lógica de relaciones de causa y efecto que no están determinadas por nadie y sólo por las cuatro fuerzas fundamentales? ¿Podemos siquiera plantearnos razonablemente tal cuestión mientras formemos parte del sistema? No otra cosa se ha discutido más o menos implícitamente en el campo del conductismo: ¿condicionamos nosotros a las ratas que estudiamos o ellas nos condicionan a nosotros? Los dueños de perros y gatos se hacen pocas ilusiones, como si sólo ellos entrenaran y condicionaran a sus animales y no al revés. Se trata de un proceso bidireccional y en el caso de los gatos es más bien unilateral, ellos tienen principalmente personal. Lo mismo ocurre con el investigador conductista y las personas "condicionadas" que se examinan. ¿Son los investigadores incondicionales sólo porque crean condiciones condicionantes? ¿El establecimiento de estas condiciones es un acto de comportamiento condicionado o una acción libre y planificada? ¿Dónde empieza y dónde acaba el condicionamiento? ¿Cómo llegamos a plantearnos estas preguntas?

Enunciados universales tan conocidos como "Todos los cretenses mienten" – presentado lógicamente por un cretense – deben, por tanto, restringirse de tal manera que los enunciados queden precisamente fuera de esta totalidad. Esto no significa otra cosa que la totalidad está limitada y no se aplica ninguna totalidad ilimitada. Esto crea espacio para algo nuevo fuera de la totalidad, para la incertidumbre. En otras palabras, el enunciado total que no puede analizarse se transforma en un enunciado relativo que, en principio, puede analizarse empíricamente. Hossenfelder no da cabida a este espacio incierto – por grande o pequeño que sea – en sus afirmaciones. Pero debería ser difícil para un científico, independientemente de su profesión, hacer afirmaciones universales atemporales sobre los seres humanos o los seres vivos en general que se formulan como leyes de la naturaleza – pero eso es exactamente lo que intenta hacer aquí. Por tanto, en su propia lógica, debe utilizar esta formulación, ya que no existe el libre albedrío. Esto significa que estas afirmaciones no dicen absolutamente nada sobre el libre albedrío, ya que la afirmación como tal no puede ser analizada. El libre albedrío, determinado por la ley natural, es por tanto de naturaleza binaria: o existe o no existe. Si no existe, todo está determinado por las partículas materiales, según Hossenfelder. El hecho de que se exprese de este modo es entonces sólo una expresión de las leyes de la naturaleza y, por supuesto, no está sujeto a su inexistente libre albedrío. Desde un punto de vista lógico – si las leyes de la naturaleza funcionan de este modo – no podemos suponer que Hossenfelder pueda cambiar sus opiniones, ya que éstas están predeterminadas por las leyes de la naturaleza. Y en caso de que cambie de opinión, esto también es sólo una expresión de las leyes de la naturaleza. Esto significa que sus afirmaciones pueden fallar en principio, pero de alguna manera no pueden ser falsificadas en la práctica, ya que se necesitarían instrumentos ajenos a las leyes de la naturaleza para verificar sus afirmaciones. Ni siquiera un largo año sabático es suficiente para resolver con éxito este problema. En resumen, esto significa que en estas condiciones prácticamente todas las afirmaciones son arbitrarias, ya que no puede haber una correlación razonable entre la realidad y las afirmaciones, ya que no podemos comprobar realmente si las cosas son como suponemos que son, porque todo está predeterminado por la ley natural a partir del Big Bang. No hay realización, ya que ésta es sólo un subproducto de los procesos físicos y, por lo tanto, no ofrece una visión real en el sentido que la gente normalmente espera y supone.

Si nos fijamos en la física como materia, se observa que la física está todavía muy lejos de lograr la tan deseada teoría de la gran unificación para combinar plenamente la física cuántica con la relatividad general/especial. Incluso es posible que se observe lo contrario, que el propio Hossenfelder (2018[a]) analiza en detalle en otro lugar, a saber, una crisis de la física fundamental. Así pues, mientras no se pueda normalizar la física, los físicos solo deberían hacer afirmaciones sobre lo que pueden hacer afirmaciones – es decir, partículas materiales – y preferiblemente no sobre fenómenos que se asignan a las ciencias de la vida en sentido amplio. Para ser justos, no es frecuente que los físicos intenten explicar el reino de la vida. Lo que falta actualmente es un modelo razonablemente bien fundado y empíricamente ampliamente probado que pueda demostrar más allá de toda duda que no hay nada fuera de la materia y que explique plenamente la vida misma, lo que incluye hacer predicciones y derivar intervenciones (véase el capítulo 1.2 (Conocimiento, comprensión y sabiduría)). No se trata aquí de demostrar que existe una dimensión (o varias dimensiones) independiente del mundo físico. Sólo que no podemos excluirla por completo. Bayesianamente hablando, asignaríamos probabilidades prioritarias a ambas posiciones tras un análisis cuidadoso. Y no parece que esta probabilidad a priori se asigne completamente a una sola posición. Incluso una probabilidad a priori mínima – ejemplo, que los procesos mentales no puedan explicarse exclusivamente por la física – conduce a ciertas dudas en la probabilidad a posteriori, y sólo eso nos basta aquí.

6.5.1.5 Conclusiones de generalización inadmisibles

Ampliando el argumento anterior, observamos en primer lugar que las conclusiones de generalización completas (= afirmaciones universales) resultan extremadamente difíciles y, de hecho, imposibles en la práctica de la investigación.

Ni siquiera el condicionamiento de Skinner dicta de forma determinista cómo discurre la curva de aprendizaje en los humanos y si discurre en absoluto según lo especificado. Lo mismo ocurre con el desarrollo del lenguaje o del pensamiento, por ejemplo. Los experimentos de Milgram sobre la obediencia

también distan mucho de ser inequívocos y generalizables, como se suele presentar en los medios de comunicación. Más bien influyen diversos factores (proximidad, aliados, presencia del experimentador, etc.).

> **Tarea 6.3: Leyes del comportamiento humano**
>
> Tarea para los lectores interesados: Piense en una ley del comportamiento humano que sea válida siempre y en todas partes, que sea válida para todos, siempre y realmente siempre en todas las circunstancias. Deje fuera las curvas de memoria de Ebbinghaus, ya que están demasiado cerca de la (neuro)biología, aunque por supuesto también tienen cierta tolerancia.
>
> Diviértete: tómate unas décadas o incluso más, porque ahí tendrás que buscar un poco.

Cuanto más cerca se mira, más difícil resulta. En lo que respecta a las ciencias de la vida (biología, psicología, pedagogía, ciencias políticas, sociología, antropología, etc.), actualmente no existen leyes universalmente válidas ni nada comparable para predecir con precisión infinita el comportamiento, los desarrollos o incluso la evolución en relación con la unidad de investigación respectiva (persona, célula, agrupación, sistema, etc.). Incluso nos atreveríamos a decir que en psicología, por ejemplo, apenas existen teorías, sino, en el mejor de los casos, planteamientos teóricos que distan mucho de cumplir los requisitos de una teoría completa y que pueden describir el comportamiento de forma estadística, pero no determinista.

Las supuestas conclusiones de generalización a menudo resultan ser no deterministas en la práctica, como – véase Hume, Popper, Lakatos, etc. – no hay ninguna garantía de que las observaciones futuras puedan describirse siempre y por completo mediante los modelos actualmente disponibles, a menos que ya lo sepamos todo sobre el mundo. Sin embargo, es probable que éste no sea todavía el caso y no está claro si lo será algún día. Recordemos que siempre podemos equivocarnos al pensar, investigar y falsar. Así que es un verdadero problema cuando la capacidad de describir partículas materiales se utiliza como base para sacar conclusiones sobre el reino de la vida sin crear las bases y demostrar que toda la vida puede reducirse exclusivamente a aspectos materiales. Las neurociencias intentan hacer algo parecido, a saber, atribuir la conciencia por completo a procesos puramente materiales en el cerebro. Esto va tan lejos que ya no se considera a la persona en su totalidad, sino sólo al cerebro (Bennett y Hacker, 2003[a]). El futuro quizá demuestre que las propias células tienen su propia conciencia y entonces la neurobiología se replanteará.

Mientras que el nivel material como al menos una base de nuestra existencia está superficialmente fuera de toda duda, esta dimensión se disuelve lentamente al examinar más de cerca la física cuántica, de modo que por debajo de la longitud de Planck ya no es posible decir lo que realmente sigue existiendo porque faltan los instrumentos. Entonces ya no se aplican las leyes anteriores, a las que se refiere Hossenfelder. Sin embargo, no hay ninguna razón para suponer que no existen más leyes de la naturaleza a este nivel sólo porque no las conozcamos. El hecho de que no podamos medir nada no significa necesariamente que esto no tenga influencia. Es cierto que tendríamos que demostrar cómo es esta influencia. A este respecto, podemos decirlo así: de momento no lo sabemos porque no disponemos de los instrumentos adecuados.

Lo mismo ocurre con uno de los modelos favoritos de los físicos en la actualidad, la teoría de strings. Ésta aún no puede probarse empíricamente y menos como modelo completo y, desde el punto de vista científico, es ciertamente interesante desde el punto de vista heurístico, increíblemente sofisticada desde el punto de vista matemático, pero en realidad carece actualmente de pruebas empíricas o incluso de aplicación. Se convertirá en relevante para la acción cuando se puedan derivar de ella procedimientos prácticos y aplicarlos con éxito y fiabilidad en la realidad o simplemente probarlos. Sin embargo, hay voces críticas que cuestionan la falsabilidad de la teoría de strings, es decir, que ni siquiera pueda estar equivocada. Además, la teoría de strings necesita más dimensiones para funcionar del todo. Nadie ha sido capaz aún de responder dónde y en qué forma existen. Los intentos de explicar esto están de nuevo más allá del empirismo. Los resultados postulados de la teoría de strings en el LHC (por ejemplo, la supersimetría) no se han materializado hasta ahora, lo que significa que la teoría de cuerdas es actualmente probablemente la teoría científica más popular

y más extendida sin base empírica. Por tanto, las cuatro fuerzas de la naturaleza mencionadas siguen dejando bastante incertidumbre. Hossenfelder no aborda el hecho de que físicos como Burkhard Heim (1925-2001) reivindicaron otras dimensiones que dan cabida a fenómenos no materiales y podrían redefinir así la cuestión del libre albedrío (von Ludwiger, 2013[a]). La teoría citada por Heim no goza de mucha resonancia en la física, es difícil de entender y actualmente no se ha comprobado empíricamente de ninguna manera. Sin embargo, esto no significa que no contenga elementos lógicamente coherentes y empíricamente comprobables. También son concebibles otras teorías físicas que podrían hacer avanzar significativamente la actual visión física del mundo. Al igual que en la discusión sobre el estudio de Bem sobre la clarividencia (véase cap. 4.4.2.2 (El estudio de Bem sobre la clarividencia)), la cuestión de si un fenómeno existe resulta ser irrelevante para nosotros aquí. Mucho más interesante es cómo se defienden metódicamente las respectivas tesis.

Así, es relevante que la argumentación de Hossenfelder contenga generalizaciones infundadas o inadmisibles que, si se examinan más detenidamente, contienen incertidumbres más que suficientes para permitir explicaciones alternativas. Por su parte, las afirmaciones generalizadas basadas en la ley están actualmente fuera de lugar. Y esa es la cuestión.

6.5.1.6 Supuestos implícitos sin base empírica de datos

La suposición de que la física puede describir completamente toda la vida simplemente porque es capaz de describir suficientemente bien el mundo de las partículas materiales es una conclusión falsa mientras no se aporten pruebas empíricas. Los datos de un acelerador de partículas o los de las órbitas planetarias o los de las galaxias, las supernovas o las estrellas de neutrones o la refracción de la luz por los cuerpos ultrapesados no sirven en absoluto para emitir juicios sobre el tema. Ninguno de ellos hace afirmaciones sobre lo vivo. Se pueden investigar aspectos materiales, pero ¿es eso suficiente? Según Hossenfelder, sí. Pero, ¿es realmente suficiente que si una ciencia no investiga algo, luego pueda hacer afirmaciones sobre ello suponiendo que son universalmente válidas?

De todos modos, es probable que a los físicos les resulte difícil investigar temas de las ciencias de la vida, ya que la mayoría de los instrumentos físicos no son adecuados para ello. Además, las partículas no se "defienden" ni desarrollan vida propia, sino que se comportan según predicciones estadísticas. Los seres vivos, en cambio, exhiben propiedades inesperadas que al menos sugieren que el *teorema de Gestalt* "El todo es más que la suma de las partes" podría implicar algo más que sólo aspectos materiales. Esto no contradice el empirismo, que demuestra que el comportamiento humano puede describirse externamente mediante modelos estadísticos. No dice nada sobre el mundo interior y los modelos estadísticos, que son válidos en grandes grupos, suelen fallar en casos individuales.

Por ejemplo, en el contexto de un experimento mental, se podría suponer que existe una dimensión espiritual separada que interactúa con la dimensión material, que puede ser descrita por la física, pero que no es 100% idéntica a ella. La materia actúa sobre el espíritu, el espíritu actúa sobre la materia, como hipótesis de trabajo inicial. El hecho de que lo vivo muera cuando muere el cuerpo no es un contraargumento aquí, ya que en principio podría ser que "algo", aquí descrito ad-hoc de forma muy poco específica como un fenómeno mente-cuerpo en el sentido de una relación causa-efecto, sobreviva como un fenómeno puramente espiritual y no dependa completamente de la materia existente y "busque" nueva materia – por ejemplo, qua una ley natural. ¿Por qué las leyes de la naturaleza sólo deben aplicarse a la materia no viva? Parece mucho más sensato suponer que existe una ley general de la naturaleza que abarca por igual el reino de la vida y el de la materia no viva. Esto iría más allá de las cuatro leyes materiales de la naturaleza que no están integradas entre sí: suficiente material para un potencial libre albedrío.

Combinemos esta idea con la idea básica del estudio de Bem (2011[a]) de que nuestro potencial humano es capaz de cosas mucho mayores de lo que se supone. Formulada en una versión que sea teóricamente sólida y pueda explicar los fenómenos parapsicológicos y al mismo tiempo vaya más allá del nivel puramente material, esto falsificaría la afirmación de Hossenfelder en el caso de un resultado positivo de un experimento replicable y serio. A saber, en el sentido de que la física no puede hacer afirmaciones definitivas sobre los seres vivos y, por tanto, la cuestión del libre albedrío sigue sin resolverse. Ésta es sólo una variante; sin duda

existen otras variantes de falsación. Básicamente, sin embargo, corresponde a la autora mostrar las pruebas correspondientes de sus tesis y la teoría no es suficiente para ello.

Estos experimentos mentales sólo pretenden mostrar que las afirmaciones de Hossenfelder son inadecuadas y que aún no se ha explorado por completo todo el espacio de hipótesis posibles, ni teórica ni empíricamente. Además, la física debería ser capaz de describir por completo y sin lugar a dudas todos los fenómenos parapsicológicos que la historia de la humanidad ha podido observar. El trabajo de Lucadou (2012[a]), físico y psicólogo que dirige el Centro de Asesoramiento Parapsicológico de Friburgo/Brisgovia e investiga seriamente los fenómenos desde una perspectiva psicológica y física, es un buen ejemplo de ello. Desde el punto de vista de Hossenfelder, todos estos fenómenos son deterministas o aleatorios, por lo que el término aleatorio es un concepto científico pobre. La coincidencia sólo engloba todos los factores que influyen y que no se pueden registrar empíricamente porque son demasiados, o que simplemente no se comprenden o no se conocen y, por tanto, no se pueden modelizar. Sin embargo, Hossenfelder debería ser capaz de hacer afirmaciones concretas sobre por qué algo resulta así y no de otra manera en casos individuales, y una referencia a las condiciones iniciales del Big Bang no es, por desgracia, suficiente.

Podría ser posible, por ejemplo, que todos los instrumentos físicos sean actualmente inadecuados para investigar una dimensión espiritual independiente potencialmente existente. Siguiendo con esta idea, se podría imaginar que una dimensión espiritual tendría que ser investigada directamente por nuestra propia mente, que podría representar un instrumento de investigación aceptable, o por instrumentos y aparatos que aún no se han inventado, para los que actualmente carecemos de tecnología. Sin embargo, lo primero requeriría presumiblemente cierto entrenamiento mental, que no todos los científicos tienen per se, ya que tales procesos de conciencia no deben confundirse con los procesos cognitivos del pensamiento (intelecto). En cualquier caso, la física resultaría bastante inadecuada para investigar tales fenómenos de forma apropiada. En tal punto, el poder explicativo de la física terminaría. El espacio de las hipótesis y el acceso a los datos empíricos están lejos de estar cubierto y, desde luego, no por la física. Sin embargo, es muy difícil discutir esto cuando alguien rechaza todas las explicaciones que no sean puramente físicas como conducentes al conocimiento. En nuestra opinión, la afirmación de Hossenfelder de una negación general del libre albedrío es una hipótesis prudente, pero de ningún modo un conocimiento empíricamente probado; y eso había que demostrarlo. Aparte de eso, la física es una ciencia comparativamente sencilla en comparación con las ciencias de la vida. Y precisamente un físico del campo de la mecánica estadística llega a esta conclusión.

6.5.1.7 Como punto (no del todo) final...

... citamos al físico E.T. Jaynes (2003[a], p.6, cursiva en el original),

> „In physics, we learn quickly that the world is too complicated for us to analyze it all at once. We can make progress only if we dissect it into little pieces and study them separately. Sometimes, we can invent a mathematical model which reproduces several features of one of these pieces, and whenever this happens we feel that progress has been made. These models are called physical theories. As knowledge advances, we are able to invent better and better models, which reproduce more and more features of the real world, more and more accurately. Nobody knows whether there is some natural end to this process, or whether it will go on indefinitely."

La lista de argumentos podría continuar indefinidamente, por ejemplo, que el libre albedrío ciertamente puede ser operacionalizado, por ejemplo a través de modelos somato-mentales, diferentes niveles del inconsciente y el condicionamiento interno de la persona que se produce, diferentes grados de la capacidad de libre albedrío, y mucho más. El uso del ejemplo del mentiroso cretino también es legítimo: si no tenemos libre albedrío, no podemos responder a esta pregunta, ya que toda respuesta está determinada robóticamente. Sin embargo, si tenemos libre albedrío, podemos responder cualquier cosa, porque sí tenemos libre albedrío y esto no significa que vayamos a responder con la verdad. Lo dejaremos en este punto; los lectores interesados pueden, por supuesto, reflexionar más. La autora Hossenfelder ha repetido su punto de vista en

posteriores entradas de blog y vídeos, pero en nuestra opinión no ha aportado nuevos argumentos sustanciales que pudieran cambiar el punto de vista.

Hay que tener en cuenta que la física no ofrece actualmente ni modelos bien fundamentados, ni datos empíricos, ni la legitimidad de una conclusión de generalización para describir y explicar la vida completamente y sin lugar a dudas. Pero este es un requisito previo para negar por completo la existencia del libre albedrío. En nuestra opinión, esto pone fin a la discusión de forma bastante definitiva con respecto a las citas pegadizas, sin que tengamos que examinar otros argumentos con más detalle. Sin embargo, muestra claramente que los motivos subjetivos, la profesión y las pretensiones de objetividad pueden desmoronarse en la práctica y dar lugar a afirmaciones y aseveraciones interesantes, e incluso directamente a creencias subjetivas. Al fin y al cabo, ¿qué hace la física como la mayoría de las ciencias? Cree en sus instrumentos, es decir, que funcionan sin lugar a dudas y son válidos. Pero ningún físico "ve" o "experimenta" directamente el principio de incertidumbre de Heisenberg, experimenta la fusión de los átomos o su desintegración o ve directamente que todo por debajo de la longitud de Planck se vuelve borroso, de modo que podemos hablar de espuma cuántica. El conocimiento científico es, por tanto, limitado y, en nuestra opinión, sólo esto basta para mostrar un poco de escepticismo ante las afirmaciones sobre el universo. Esto se aplica no sólo a Hossenfelder, sino en general.

Las dificultades son, por tanto, de naturaleza múltiple y ya son visibles antes incluso de que el concepto de libre albedrío entre en escena. Esto significa que nuestro tema de la objetividad – es decir, la pretensión de Hossenfelder de una pretensión general de explicar el comportamiento vivo – puede ser completamente exagerado y en su lugar termina siendo completamente subjetivo, ya que la autora tiene obviamente una fuerte necesidad de hacer afirmaciones arrolladoras, generalizadoras y atemporales sobre el libre albedrío y no acepta ninguna otra posición, al menos no como posible. Ella cree en sus teorías, lo cual no es un problema en sí mismo, pero no debemos confundir esto con un conocimiento fiable. Esto también se aplica a todas las ciencias. Recordemos a Kuhn (1973[a]) y Feyerabend (1976[a]), que mostraron cómo cambia socialmente el conocimiento reconocido y qué puede ser el conocimiento en absoluto. En un intento de ser especialmente objetivo y generalizador, Hossenfelder acaba en un dilema subjetivo de los "cretenses mentirosos" desde nuestro punto de vista. Puesto que, desde su punto de vista, no hay libre albedrío, no expresa la opinión que ha elegido, sino sólo lo que le dictan los procesos físicos en su interior. Ni siquiera se puede hablar de razonamiento, porque éste presupone un pensamiento libre activo. ¿Es posible discutir o incluso juzgar algo así? No desde su punto de vista, porque no tiene libre albedrío que expresar. Puesto que la física determina sus acciones, está completamente atrapada en su subjetividad, porque por definición no puede hacer otra cosa y percibirla de la mejor manera posible, e incluso eso está determinado. Ya no es objetiva. La objetividad sería una descripción correcta e independiente de una situación o hecho, más un análisis, etc., pero esto presupone una cierta libertad de acción en alguna parte para poder lograrlo en absoluto. De lo contrario, el término carece de sentido. La autoaplicación de Hossenfelder a Hossenfelder significa, por tanto, que sus tesis pierden toda pretensión de objetividad y se vuelven literalmente vacías, quedando sólo ella como sujeto en su punto de vista. Esta sería la aplicación consecuente de sus tesis. No hay nada más aplicable a nuestro texto. Por supuesto, ya ha sido definido por el Big Bang y lo estamos escribiendo sobre la base de estrictos procesos físicos de causa y efecto. Si permitimos una visión normal del mundo, Hossenfelder también nos parece atrapada en su subjetividad, ya que no deja espacio para el "*también podría ser diferente*".

Volviendo a la objetividad, ésta a menudo se operativiza y se comunica de tal forma que un proceso tiene lugar, en la medida de lo posible, sin el ser humano, y la subjetividad implica explícitamente al ser humano y, por lo general, a su capacidad de reflexión, lo que no excluye condicionamientos y puntos ciegos de cualquier tipo. Por desgracia, ambas posturas pueden exagerarse. Lo subjetivo se define a menudo por lo contrario de lo objetivo y viceversa, lo que en sí mismo no favorece mucho el conocimiento y recuerda más bien a la *no falsabilidad* del modelo freudiano de psicoanálisis: el id explica el superego y el ego, el ego explica los otros dos y el superego es tratado de la misma manera. Por ejemplo, las directrices de un procedimiento objetivo siempre están hechas por el hombre y, por lo tanto, esto también se mueve en el ámbito de la subjetividad, porque alguien lo inventó en algún momento, elaboró directrices y reglas, etc. Incluso los algoritmos y las matemáticas no son objetivos en el sentido de que se han creado a sí mismos a partir de la realidad. En los cálculos estadísticos intervienen muchos elementos subjetivos y no sólo en la cuestión de fijar umbrales críticos de superación; no en vano hay revistas que sólo se ocupan de comparar

procedimientos y algoritmos y los respectivos representantes defienden con vehemencia sus posiciones. Si todo esto fuera objetivo e incuestionablemente natural, nada de ello sería necesario. El extremo en la subjetividad comienza cuando las afirmaciones pierden su carácter científico y su conexión con el discurso general, y no nos referimos al tipo de discurso en el que todo el mundo está de acuerdo con las tesis que uno mismo sostiene de todos modos. En nuestra opinión, la única forma de integrar objetividad y subjetividad es mediante la intersubjetividad, es decir, el discurso crítico y el diálogo con los demás (Bohm, 1998[a]). Así pues, si se quieren formular enunciados libres de contexto y semejantes a leyes, el camino para llegar a ellos debe estar libre de dudas. Esto no niega lo que la física ya puede explicar o modelar. Simplemente afirma que el mundo entero aún no está explicado y que existe suficiente incertidumbre residual para un gran número de fenómenos. El origen de la vida sería uno de ellos. Hasta la fecha, ningún científico ha sido capaz de explicar completamente el mundo.

También es apasionante que la lógica aristotélica del *si-entonces* o del o *bien-o bien* opere generalmente en el trasfondo de estos argumentos. Hay otras formas de razonamiento, como la antiquísima lógica del tetralema, que funciona más según el principio de "*lo uno y lo otro*". Tiene los elementos "*lo uno, lo otro, ambos, ninguno, algo completamente diferente*". La dialéctica hegeliana, que intenta integrar posiciones obviamente contradictorias en un nivel superior de cognición y las identifica como formas compatibles de expresar una misma cosa, también sería pragmáticamente útil, pero en última instancia representa un subconjunto de la lógica tetralemática. Cabría preguntarse aquí qué ocurriría con el debate subjetividad-objetividad si simplemente cambiáramos la lógica subyacente. ¿Podrían simplemente mantenerse las posiciones a menudo endurecidas descritas hasta ahora? ¿Seguirían siendo tan opuestas y mutuamente excluyentes? Esa sería una tarea sobre la que los lectores deberían reflexionar.

Como digresión final: ¿Cómo tratar con uno mismo y protegerse y proteger a los demás de las proyecciones mentales y las profecías autocumplidas para evitar cometer los errores comentados anteriormente?

> Tarea 6.4: Protección frente a uno mismo
>
> - No te limites a buscar la confirmación positiva de tus propias suposiciones, sino que investiga sistemáticamente lo contrario de tus propias suposiciones en el sentido de Popper. Así no te tomarás demasiado en serio ni te sobrevalorarás.
>
> - Comunícate regularmente con personas que no tengan tu misma opinión o – mejor aún – que sostengan lo contrario de tus propias suposiciones. Deberías aprender de ellos, tomarlos en serio y probar sus puntos de vista en la práctica para ver si, después de todo, podrían ser legítimos.
>
> - Reflexione sobre sus propias acciones no sólo en el ámbito cognitivo, sino también en el mental y emocional.

Esto último es una de las cosas más difíciles de hacer, que en contextos profesionales se da más entre psicoterapeutas y poco o nada entre científicos.

6.5.1.8 Volviendo a la teoría de Bayes

¿Cómo interpretar esta larga digresión a lo largo del teorema de Bayes? Carecemos de un estudio exhaustivo sobre la probabilidad de una dimensión mental aparte de la física, que junto con ésta pueda explicar la vida. Pero los argumentos no han sugerido en absoluto que esta dimensión no pueda existir en todas las circunstancias. Una dimensión espiritual más allá de la pura física bastaría para falsear las afirmaciones de

Hossenfelder. Sin embargo, no podemos fijar la probabilidad prior de que exista de forma generalizada, sólo porque nos guste o porque creamos que podemos hacerlo fácilmente. Por otro lado, podemos descartar la posibilidad de que ya se haya hecho todo lo posible para excluir por completo y para siempre esta dimensión en cualquiera de sus formas. Es decir, que exista una probabilidad prior de una dimensión espiritual mayor que cero, pero con un nivel muy incierto. Esto deja automáticamente espacio para que el libre albedrío, en el sentido de una probabilidad posterior mayor que cero, exista en principio, aunque la probabilidad pueda ser baja. Una probabilidad prior muy baja combinada con un experimento débil y pequeño conduce a una probabilidad posterior muy baja. Sin embargo, esto es suficiente para terminar esta discusión, porque no queremos demostrar nada aquí, excepto que Hossenfelder ha trabajado incorrectamente, estadísticamente hablando. No investigó seriamente todo el espacio de posibilidades, sino que se quedó en la física, lo que no es suficiente según los puntos anteriores. Debería haber quedado claro lo rápido que se difunde la propia opinión en ciencia bajo la apariencia de objetividad, es decir, se actúa subjetivamente. Es difícil argumentar en contra de esto.

Aquí, en el ejemplo, no queremos afirmar lo contrario, que existe el libre albedrío y que siempre está disponible (o no), sino sólo subrayar lo abierto de la cuestión. Para ello es necesario (por ejemplo, en el contexto del análisis cualitativo de datos) reflexionar sobre el uso de las propias creencias subjetivas. En nuestro caso, éstas serían que operacionalizamos un libre albedrío como la capacidad de los seres humanos de hacer conscientes en principio sus propias tendencias inconscientes y dejar entonces de comportarse según condicionamientos inconscientes. Esto corresponde a la transición del comportamiento al hacer y al actuar. La acción psicológica, la educación y la psicoterapia implican más o menos implícitamente el concepto de libre albedrío. Sin embargo, empíricamente sospecharíamos que en la vida cotidiana la mayoría de las veces no hacemos uso de nuestro potencial de libre albedrío, por lo que a menudo parece como si no existiera. Si no existiera el libre albedrío, Hossenfelder se vería obligada a hacer sus afirmaciones. Estaríamos obligados a argumentar como tenemos que hacerlo y la situación en sí no estaría abierta a discusión, ya que cada uno sólo diría lo que le dicta su inexistente libre albedrío. Entonces, Hossenfelder no llega a sus afirmaciones por voluntad propia, sino que son producto de las fuerzas de la naturaleza, y también de nuestras reacciones ante ellas, etc. Podríamos seguir con esta argumentación hasta ponernos morados. A partir de ahí, se puede ver cómo funcionan todas las afirmaciones.

Sin embargo, la realidad no se corresponde con la inducción matemática. Continuemos con las creencias subjetivas en el contexto de la estadística de Bayes.

6.5.2 Creencias subjetivas en la estadística de Bayes

La larga discusión que sigue sobre objetividad y subjetividad es útil porque saca a la luz supuestos implícitos fundamentales de la ciencia. En realidad, la diferenciación entre subjetivo y objetivo resulta obsoleta, ya que éste no es el problema central. El problema central de la ciencia es: *¿Qué hacemos de una situación, cómo desenterramos el conocimiento real?* Que la cognición provenga de la pura reflexión subjetiva, de la observación, de la experimentación o de todo lo anterior, mientras pase un control crítico de la realidad, a la larga es irrelevante. Recordemos un acontecimiento científico revolucionario: el descubrimiento del anillo bencénico por August Kekulé (1829-1896) se lo debemos a un sueño y a su creatividad, así como a bastante trabajo previo, es decir, a algunos acontecimientos altamente subjetivos, así como objetivamente observables. Sin embargo, Kekulé fue capaz de extraer de ellos deducciones científicas que son transparentemente reproducibles; de lo contrario, la química moderna no podría funcionar como lo hace. Y lo hace hoy en día siguiendo la lógica y la experimentación científica. Lo mismo podría decirse de las teorías de Freud, Piaget, Skinner y muchos otros: estudios aislados, prácticamente sin estadísticas o con estadísticas mínimas o incluso, como la teoría de Freud, en realidad no falsables en absoluto. Sin embargo, son la fuente de hallazgos exhaustivos que siguen llenando los libros de texto.

El propio concepto de objetividad no es – como cabría esperar – objetivo, del mismo modo que la diferencia entre significativo y no significativo no es en sí misma significativa (Gelman & Stern, 2006). Algo se considera objetivo si puede demostrarse independientemente de las condiciones y los observadores. Sin

embargo, esto suele aplicarse sólo a hechos abstractos, como los de las matemáticas. Pero incluso un ejemplo como 1 + 1 = 2 sólo es objetivo en el sentido de que las reglas subyacentes (adición, números en sí, ...) se reconocen axiomáticamente como convención. En las ciencias empíricas que se ocupan de cuestiones muy complejas, incluidas todas las ciencias sociales, el concepto de objetividad está parcialmente justificado en los laboratorios, pero rara vez en los estudios de campo o ante modelos complejos con muchos parámetros seleccionados. Por ejemplo, Hume, partiendo del supuesto de que nuestros sentidos son la única fuente de información sobre el mundo exterior, se preguntó: "¿Qué es el mundo exterior? ¿Existe independientemente de nuestros sentidos y de la percepción?

También es interesante aquí el punto de vista de Popper (1943), defensor de la objetividad, quien suponía que si las suposiciones sobre el mundo son tan actuales, esto lleva a que sean verdaderas aunque nadie las observe. Por otra parte, puede rebatirse que esto está por definición fuera de discusión, ya que tal ejemplo (abstracto) sólo es posible en un experimento mental. En consecuencia, nos preguntamos: ¿ha hecho Popper alguna vez una investigación empírica? Son precisamente los efectos cuánticos como el entrelazamiento o el experimento de la doble rendija los que muestran la dependencia de la medición de la observación y su intención (es decir, el ajuste) y el vínculo inseparable de los resultados físicos de la medición de la disposición instrumental y la observación humana. Incluso un instrumento de medida que mida, evalúe, calcule, etc. automáticamente sigue necesitando un ser humano para interpretar el contenido, que o bien lee el resultado o bien programa algo que haga lo mismo. El principio de incertidumbre de Heisenberg (= imposibilidad de determinar simultáneamente la ubicación y el momento de una partícula al mismo tiempo) resume la conexión entre la observación y el efecto sobre lo observado. Sin embargo, no hay que concluir de ello que exista distorsión alguna. Más bien, el fenómeno es la determinación de un punto de vista (partícula, onda) desde el que se observa todo lo demás en el mundo.

Estamos de acuerdo con Popper en que demostrar las teorías y ponerlas a prueba de forma crítica puede llevar a la ciencia a un nivel tan alto que las dudas en la teoría y el empirismo asociado se vuelven menores. También ayuda examinar hechos sencillos, como intenta hacer la física al reconstruir las leyes de la naturaleza. En las ciencias sociales, las cosas son probablemente mucho más complejas. Aunque el comportamiento de los organismos vivos puede describirse en principio con las mismas matemáticas y los mismos modelos matemáticos que los procesos físicos cuánticos, es obvio que las interacciones cuánticas en los seres humanos o entre muchos seres humanos son mucho más complejas que las investigaciones de "un puñado" de partículas en un acelerador de partículas. Y ni siquiera hemos añadido la dimensión de lo vivo y su imprevisibilidad.

Pero dado que la objetividad está determinada por los humanos, tanto a través de los datos y su interpretación como a través del metadiscurso filosófico, la objetividad generalmente resulta en una operacionalización que en la práctica científica (metateoría así como empirismo) simplemente significa intersubjetividad – el *estado actual y discursivo del error*. Puesto que somos sistemas neurobiológicamente cerrados (Maturana & Varela, 1984) y no podemos percibir nada que esté fuera de nosotros, sino que sólo podemos trabajar con lo que nos ofrecen nuestros sentidos, la cuestión de la objetividad completa es trivial. No existe, y si existiera, nosotros como sujetos ya no podríamos percibirla, puesto que ya la estamos procesando y distorsionando a través de nuestra percepción. Así, el biólogo von Uexküll (1920, p.338s.) llama objetividad a una "conveniencia del pensamiento" que surge de convenciones. Lo que sería necesario para una verdadera objetividad es una percepción del mundo en la que nosotros mismos estemos física y mentalmente "recortados" – una idea entre interesante y extrañamente morbosa, pero en cualquier caso imposible. En el plano político, no sólo hemos visto recientemente (palabra clave: "hechos alternativos", que no son más que creencias no demostradas) cómo se reinterpreta la objetividad o incluso la intersubjetividad para imponer los propios intereses (de poder), practicar el populismo y, en un sentido más amplio, llevar a cabo un lavado de cerebro. No sólo existe la inteligencia de enjambre y la intersubjetividad por sí sola no es garantía de calidad. Así, el concepto de objetividad adquiere una nueva legitimidad desde el punto de vista estratégico para ejercer influencia en el mundo. Weber (1904/1991) sostiene aquí que la cognición tiene lugar en función de ideas de valor y, por tanto, está específicamente situada en cada caso (véase también Stegmüller, 1973a, cap. 1.IX sobre el postulado de la libertad de valor, según el cual los enunciados no

pueden ser tratados como verdaderos o falsos simplemente porque a uno subjetivamente le gusten o no). Dentro de la estadística de Bayes se pueden determinar a grandes rasgos tres corrientes

- la de las creencias *subjetivas*,
- la de la orientación *(más) objetiva*, centrada en la comprobación empírica crítica, y
- la orientación *radicalmente objetiva* (Bayes empírico).

Las corrientes difieren en la modelización del concepto de probabilidad y la correspondiente elección de la Prior.

6.5.2.1 Creencias subjetivas

La dirección subjetiva sostiene que todas las conclusiones son siempre relativas a un determinado nivel de conocimiento. Este nivel de conocimiento varía de un individuo a otro. Se asume que un conocimiento objetivo de la realidad física es un ideal inalcanzable. De Finetti (1974) lo exagera cuando escribe que la probabilidad no existe y (presumiblemente) quiere decir que la probabilidad es una construcción hecha por el hombre y no existe independientemente de nosotros (véase también Nau, 2001). La aplicación de la probabilidad a priori es, por tanto, una cuantificación de las creencias personales dentro de un contexto, que se contextualizan con los datos mediante el teorema de Bayes. Desde un punto de vista epistemológico, se podría argumentar que la probabilidad es una *puntuación* en el sentido de Bateson (1985) de hacer el mundo manejable y comunicable y, en esa medida, equivale a una función modelo. Esto tiene poco que ver con la realidad. No es de extrañar que tal punto de vista suscitara críticas y, en particular, cuestionara la cientificidad de la estadística bayesiana, ya que aquí supuestamente prevalece el deseo en lugar de la objetividad. En cuanto a la elección de las distribuciones a priori, parece que los científicos pueden servirse como en una cacharrería y coger lo que más les convenga subjetivamente. Como contraargumentos a esta opinión, se pueden esgrimir los argumentos ya mencionados de que la objetividad no está mucho más claramente definida que la subjetividad. La elección de una prior debe hacerse siempre de forma clara y transparente. Como se ha mencionado varias veces, este requisito no lo cumple necesariamente la estadística frecuentista y las decisiones respectivas en el contexto de estudios concretos, especialmente la justificación de los límites críticos de decisión (Wagenmakers, 2007c). Si sustituimos *subjetivo* por *intersubjetivamente comprensible* a lo largo de criterios transparentes en un contexto de máxima incertidumbre y dados más o menos datos empíricos, la situación cambia de repente. En resumen, el uso del término subjetivo induce a confusión y, por lo tanto, debería sustituirse por una elección argumentativa de lo anterior según criterios científicos comúnmente aceptados, si bien somos conscientes de que el propio término "aceptado" puede ser un problema.

6.5.2.2 Elección objetiva

La otra dirección (más) objetiva elige la Prior – porque éste es el eje de toda la discusión sobre objetividad y subjetividad – de acuerdo con reglas preestablecidas (Kass & Wasserman, 1996). Entre ellas se incluyen enfoques como el de Jeffreys (1939/1961), de modo que la Prior es invariante a ciertas transformaciones, o la elección según el principio de máxima entropía (véase el capítulo 6.14, Jaynes, 1957a, 1957b, 1986b). Lo que tienen en común las Priores objetivas es que son bastante poco informativas y a menudo representan un estado ingenuo del conocimiento. Esto es obviamente diferente de la postura subjetiva, en la que las Priores reflejan el grado de convicción personal. Ahí radica una modelización diferente del concepto de probabilidad. La noción de elegir una Prior siguiendo reglas explicables, de modo que no sea de naturaleza arbitraria sino que represente el estado actual del conocimiento, es algo con lo que sólo se puede estar de acuerdo. Desgraciadamente, en la práctica esto puede llevar a defender el uso de tales reglas como una norma siempre estándar, lo que a su vez es lamentable. Básicamente, la dirección objetiva en la estadística de Bayes supone que, dada la misma información a priori y las distribuciones a priori derivadas de ella,

diferentes investigadores deberían llegar exactamente a los mismos resultados respecto a las afirmaciones a posteriori. Los resultados no se entienden como un grado de creencia subjetiva, sino como el resultado de un proceso transparente y razonado para aprender coherentemente de los datos en el marco de la metodología bayesiana. Esto no excluye incorporar información contextual, como se discute ampliamente en Jaynes (2003), por ejemplo.

Sin embargo, la forma de integración debe ser intersubjetivamente (= objetivamente) comprensible. No se trata de opiniones, sino de información verificable. La objetividad también puede ser exagerada si, por ejemplo, la información contextual no se transfiere a una Prior, sino que se elige en función de consideraciones matemáticas – cuestionable, aunque sea matemáticamente correcta, elegante, etc.

Volvamos al sentido común, que Caticha (2009, p.3, cursiva en el original) recomienda:

> „Our objective is neither to assess nor to describe the subjective beliefs of any particular individual. [...] Our concern here is not so much with beliefs as they actually are, but rather, with beliefs as they ought to be. Rational beliefs are constrained beliefs. Indeed, *the essence of rationality lies precisely in the existence of some constraints*. The problem, of course, is to figure out what those constraints might be.We need to identify normative criteria of rationality. [...] Here is our first criterion of rationality: whatever guidelines we pick they must be of general applicability — otherwise they fail when most needed, namely, when not much is known about a problem."

Gelman (2005) ofrece un punto de vista pragmático sobre la práctica concreta de la investigación en relación con el punto de vista objetivo frente al subjetivo.

6.5.2.3 Bayes empírico

Sin embargo, existe otro subtipo de la dirección objetiva, a saber, los representantes de los enfoques empíricos de Bayes (por ejemplo, Casella, 1985). En este caso, la variable Prior se estima directamente a partir de los datos empíricos disponibles. Habría que examinar si esto tiene sentido y no es más bien redundante y conduce luego a profecías autocumplidas. Si dejamos de lado las matemáticas, la estimación de la prior a partir de los datos no significa otra cosa que, en primer lugar, descuidar por completo toda la demás información situacional-contextual. Además, este procedimiento contradice el procedimiento clásico de Bayes, que espera una Prior fija antes de fijar los datos. Además, a los datos – que pueden ser aleatorios – se les asigna demasiada importancia y parece como si el proceso de autoactualización del aprendizaje a partir de la experiencia experimentara un cierto estancamiento. No hay que olvidar que los datos se utilizan dos veces: una para la probabilidad a priori y otra para la probabilidad probable, si se quiere aplicar el teorema de Bayes de forma exhaustiva. Con cierta malicia, se podría insinuar aquí una falta de sentido común.

Existe el peligro de descuidar el pensamiento y preferir los análisis (semi)automáticos (Gelman, 2008a; Gigerenzer & Marewski, 2015), que rara vez resultan especialmente rentables a largo plazo. Sin embargo, si se deja de pensar adecuadamente en lo previo, existe el peligro de que, dada la potencia informática disponible, el sentido común se limite inmediatamente a unos pocos algoritmos y datos. En la estadística clásica existe el requisito de que la significación la determine el ser humano y no la máquina o el programa informático, lo que desgraciadamente no significa que así se haga en la práctica. Este requisito requisito es aparentemente necesario en el futuro en la estadística de Bayes (véanse las explicaciones en este libro sobre los factores de Bayes). Sin embargo, existe otra variante de interpretación, a saber, que Empirical Bayes es una aproximación de la estimación bayesiana común en el marco de un modelo jerárquico, de modo que los parámetros del nivel más alto se fijan en sus valores más plausibles recién estimados a partir de los datos en lugar de integrarlos. Ahora bien, ¿por qué este esfuerzo supone una ventaja con respecto al procedimiento Bayes convencional, que en primer lugar utiliza toda la información y en segundo lugar es accesible a cambios y nueva información sigue abierto.

Wagenmakers (2007c), representante de la dirección objetiva de Bayes y partidario de los factores de Bayes, discute a lo largo de la discusión objetiva frente a subjetiva utilizando varios ejemplos hasta qué punto

la prior influye o no en la posterior. Si examinamos más detenidamente el teorema de Bayes (véase el capítulo 6.4), observaremos que en el caso de un término producto, un elemento no puede simplemente ignorarse. Si añadimos explícitamente la importancia de la información previa y comprendemos que el teorema de Bayes permite la posibilidad de aprender de la experiencia – es decir, que los nuevos conocimientos se convierten en nuevos conocimientos previos, que a su vez pueden combinarse con nuevos datos, y que esto puede practicarse durante mucho tiempo como un proceso iterativo – resulta sorprendente la idea de que la información previa debería tratarse como *ingenua*. Esto equivaldría a un imperativo de "olvida tu conocimiento acumulado sobre este contexto y finge que no has oído hablar de él". Esto va en contra de la lógica de la ciencia, aprender del conocimiento existente y hacerlo mejor en el futuro. Si damos la vuelta al argumento, podemos formular polémicamente y demandar de los estadísticos clásicos y a los estadísticos de Bayes – que ignoran el conocimiento cualitativo para la Prior -- de utilizar por fin el conocimiento previo razonado y no hacerse parecer más estúpidos de lo que son. La lógica estructural científica también exige que los resultados de las inferencias sean los mismos e independientemente de que primero sirva de base el Prior #1 y después la Prior #2 o viceversa, siempre y cuando los datos sean independientes entre sí y no se hagan pruebas donde ya se ha desarrollado exploratoriamente el modelo. Como se explica en otro lugar (palabra clave: regularización, véase el capítulo 6.12), la elección de la prior permite restringir el intervalo válido de valores. En concreto, los valores que no son plausibles en la realidad pueden excluirse o su influencia puede minimizarse, si se trata de un conocimiento aceptado y justificado. Entonces se excluyen estos valores y se descartan todas las conclusiones posteriores sobre ellos. La exclusión explícita de valores puede tener un efecto desfavorable. Alternativamente, uno puede modelar ciertos valores como muy inverosímiles vía la prior.

Rouder, Lu, Speckman, Sun y Jiang (2005) ponen el ejemplo de un partido de béisbol en el que un no favorito va ganando 5-0 al favorito después del primer *inning*. Según la máxima Likelihood, en ese momento se podría concluir que con 9 innings el marcador sería 45:0 al final. Sin embargo, tal valor es tan poco realista que la selección de una prior basada en este innings puede llegar a la conclusión de que es posible que los no favoritos no caigan del todo (¡a menos que conozcamos el discurso del entrenador en el vestuario!), pero el éxito después del 9º inning sigue siendo bastante improbable. Por lo tanto, la información previa debe utilizarse de forma justificada y no a ciegas. Studer (1996b) argumenta de la misma manera sobre la cuestión de si el éxito es posible en la terapia de la adicción. Sabemos por el pasado que sí, que es posible, pero ni al 0% ni al 100%. Si los tres primeros clientes tienen un éxito del 0 %, no caemos en la depresión, y a la inversa, no nos dejamos llevar por la euforia si los tres primeros clientes tienen un 100% de éxito. Sabemos que ambos extremos son posibles y no sabemos en qué secuencias se manifiestan. No obstante, calculamos probabilidades posteriores de éxito, pero asumiendo una estimación robusta para que las tolerancias reflejen nuestras hipótesis. En este caso, nuestras hipótesis dicen que tanto el éxito como el fracaso son posibles y no pueden excluirse como extremos. En la práctica, esto significa que después de los primeros clientes con un 0%, aún inferimos una eficacia ligeramente superior al 0%, y después de los primeros clientes con un 100%, inferimos una eficacia ligeramente inferior al 100%.

Los enfoques bayesianos "objetivos", como los propugnados por Berger (2006) y otros (p. ej.), parecen elegir las priores en función de consideraciones matemáticas y de acuerdo con propiedades particulares (p. ej., invarianza a las transformaciones, uniforme, prior de Jeffrey, prior de máxima entropía, prior de referencia, prior débilmente informada, etc.) en lugar de basarse en priores cualitativos de contenido, pero a veces en relación con la Likelihood (Gelman, Simpson & Betancourt, 2017). Esto conduce a una preferencia general hacia las priores vagas o no informativas y, en términos de contenido, a descuidar la información contextual disponible.

Por lo tanto, la objetividad se operativiza a menudo a través de características matemáticamente deseables, que nos parecen interesantes y problemáticas. Interesante, porque puede facilitar los cálculos; problemático, porque la idea de conocimiento previo se lleva así ad absurdum, salvo en el caso de que falte conocimiento cualitativo previo. Si no existe conocimiento previo, uno puede, por supuesto, tomar el grado de convicción subjetiva para crear una prior – o mantenerlo tan vago y desinformado como sea posible. Ambas opciones parecen justificables. Parece difícil situar las matemáticas por encima del contenido, porque entonces es más fácil calcular, pero quizá se excluyan cosas significativas. Sólo cuando el contenido está fijado, debe permitirse que las matemáticas introduzcan trucos, pero no a expensas de la información cualitativa.

En este punto recordamos una y otra vez que los modelos son fundamentalmente sólo modelos. Esto es cierto para todos los modelos a priori. Diaconis y Freedman (1986, p.11) resumen la situación de la siguiente manera:

> „It is useful to separate Bayesians into two groups: we will call them 'classical' and 'subjectivist.' Classical Bayesians, like Laplace or Bayes himself, seemed to believe there is a true but unknown parameter which is to be estimated from data. This parameter is part of an objective probability model for the data. Prior opinion about the parameter is expressed as a probability distribution. Subjective Bayesians like de Finetti and Savage reject such ideas; for them, probabilities represent degrees of belief and there are no objective probability models."

Así, además del diseño general de un estudio, la elección de los parámetros del modelo nunca puede considerarse objetiva o independiente del investigador. En el enfoque bayesiano objetivo, la elección de la información previa se realiza mediante un procedimiento basado en convenciones, es decir, según reglas predefinidas. Sin embargo, esto no garantiza la objetividad tal y como algunos la entienden.

Un concepto de objetividad con el que estamos plenamente de acuerdo lo resume Wagenmakers (2007c, p.2, cursiva en el original) sobre el trasfondo de argumentos similares a los aquí recogidos, "Así pues, considero que la objetividad significa que, dados los mismos datos *y* los mismos supuestos relativos al modelo, diferentes investigadores llegarán a las mismas conclusiones". Esto permite la intersubjetividad, el debate y el diálogo, características de la buena ciencia. Es útil exigir que la investigación no dependa (exclusivamente) de la persona del investigador, un argumento que también necesita unas discusiones en análisis cualitativos. Como siempre, una salida parece estar en el discurso crítico y la transparencia para replicar y revisar críticamente los análisis. Existen diferentes enfoques para la construcción objetiva de distribuciones a priori (Wagenmakers, 2007c, p.2):

- La Prior contiene tanta información como una sola observación ("Prior de información unitaria", Kass & Wasserman, 1996),
- invarianza tras la transformación de los datos (Jeffreys, 1939/1961) y
- maximización de la entropía (Jaynes, 1968).

Por un lado, un enfoque basado en reglas tiene la ventaja de respaldar el concepto de objetividad antes mencionado y también la exigencia de discurso y transparencia. Con muchos parámetros en un modelo, también es pragmático y eficaz, ya que se aplica la misma lógica a todos los parámetros. Por otro lado, existe el riesgo de que los resultados no se incluyan, lo que supone una pérdida de información. Si esto ocurre teóricamente y prácticamente hasta tal punto que realmente haya que hablar de una pérdida sigue siendo una cuestión difícil y quizás en casos concretos incluso irresoluble. Por desgracia, no existen criterios claros para responder satisfactoriamente a este tipo de ambigüedades, ni siquiera en el caso de preguntas más sencillas. Sin embargo, la influencia de la respuesta a priori varía en función del modelo y el método de análisis. Dos casos de ejemplo ilustran las posibles diferencias:

1. La estimación de un valor medio a partir de datos distribuidos normalmente. Si se toma aquí una Prior no informativa, como una distribución uniforme que no favorece un rango de valores, los datos (= Likelihood) determinan la Posterior con bastante rapidez, es decir, desde el principio. Esto demuestra rápidamente su robustez frente a distintas Priores a medida que aumenta la cantidad de datos. El estudio de caso sobre los porcentajes de aprobados en la terapia de la adicción (véanse los capítulos 6.15.2 y 6.15.2.1) ilustra dicha robustez, ya que frente a la Prior, la totalidad de los datos determina la Posterior. Si se repite el teorema de Bayes continuamente con cada nuevo aumento de datos (por ejemplo, las tasas de aprobados de un año a otro) e incluye la Posterior $_{t-1}$ anterior como la Prior $_t$ (=aprender de la experiencia) la Prior $_t$ (= Posterior $_{t-1}$ de antes) determina la nueva Posterior$_t$ y los datos recién recogidos (= Likelihood) sólo tienen una influencia modesta. Siempre depende de la perspectiva que se adopte. En modelos más complejos con diferentes supuestos de distribución para los parámetros, esta robustez ya no es tan fácil de conseguir o se requieren tamaños de muestra mucho mayores, para que la probabilidad

compense las expectativas previas desfavorables. En casos extremos, incluso esto sólo puede tener éxito hasta cierto punto, por ejemplo si valores absurdos o casi imposibles determinan la prior y no son en absoluto plausibles empíricamente, es decir, no se dan.

2. La situación es diferente con los factores de Bayes. Según Lindley (1957) y Shafer (1982), una hipótesis bilateral es sensible a la Prior. Esto puede entenderse fácilmente (Wagenmakers, 2007c, p.2) en el sentido de que si se aumenta el rango permitido de valores para una hipótesis, aumenta al mismo tiempo la complejidad de la hipótesis. De este modo, el procedimiento abarca valores más improbables en la hipótesis a priori y disminuye la Likelihood media de los datos observados dada dicha hipótesis. Dado que el BF21 (= factor de Bayes, véase más adelante el capítulo 6.8.1 sobre la prueba de hipótesis de Bayes) es un cociente entre dos probabilidades y el cambio en las expectativas después de ver los datos, este cociente cambia en consecuencia. Merece la pena examinar el ejemplo con más detalle, ya que la lógica subyacente y los problemas asociados a ella pueden verse en una cita de Wagenmakers (ibíd.) en el mismo punto:

> "Para un bayesiano subjetivo, esto no es realmente un problema, ya que *Pr(*μ*)* refleja su creencia previa. Para un bayesiano objetivo, la comprobación de hipótesis constituye un reto mayor: por otro lado, una prior demasiado vaga no es realmente un problema. Por otro lado, una prior demasiado vaga puede aumentar la complejidad de H_1 hasta tal punto que H_1 siempre tendrá una probabilidad posterior baja, indepen-dientemente de los datos observados".

6.5.2.4 Conclusión intermedia: ¿subjetiva u objetiva?

Realmente no podemos estar de acuerdo con ninguna posición. No la "subjetiva" en el sentido de que no se trata del grado de convicción personal en la ciencia. En el ámbito de la ética y la aceptación de responsabilidades sería el término adecuado, pero aquí se trata de argumentos anclados y fundamentados en la materia que son susceptibles de discurso. Y las convicciones subjetivas no representan un estado de conocimiento, sino una actitud personal. La pregunta es: "¿Existe o no el conocimiento previo?". Si existe y no se utiliza, no es científico. Si no existe, hay que modelizar adecuadamente el estado de no-conocimiento, lo que es posible con Máxima Entropía (véase el capítulo 6.14). A la inversa, si el conocimiento previo indica que una prior es imprecisa y, por lo tanto, aumenta la complejidad de un modelo, esto es una expresión del estado de error actual y discursivo. Ignorar esto, aunque no haya un conocimiento mejor, parece tan poco científico como asignar una prior según el buen o mal humor de la mañana. Esta supuesta discrepancia tal vez podría resolverse si finalmente ya no hubiera distinción entre subjetivo y objetivo, sino un acuerdo sobre el uso del conocimiento previo existente, verificable y discursivo. Éste puede, como ya se ha explicado, surgir de la visión interior subjetiva de la vida cotidiana, pero debe, no obstante, traducirse a una forma científica para que sea verificable críticamente. Un procedimiento qua convención es siempre problemático, aunque la intención sea correcta. La otra cara de la moneda suele ser que hay demasiados que se limitan a adoptar estas convenciones irreflexivamente y qua norma y no se cuestionan si tiene algún sentido en el caso concreto.

En la estadística clásica hemos visto lo que ocurre cuando se impone la convención en lugar de la orientación al problema y la adecuación al caso. Sigue habiendo suficientes decisiones subjetivas a lo largo del proceso de investigación (diseño, instrumentos de encuesta, muestras, etc.) y también dentro del análisis estadístico de datos (por ejemplo, elección del procedimiento de análisis de datos, decisiones sobre los parámetros a lo largo del ajuste del modelo, etc.) que resquebrajan bastante o incluso socavan el predicado objetivo en el sentido de independencia del investigador. Sería fundamental preguntarse: ¿para qué sirve realmente la objetividad? La estandarización socava fácilmente la creatividad y la responsabilidad personal por las propias acciones.

La objetividad total eliminaría al investigador del proceso y ¿qué quedaría? ¿Acaso las preguntas de investigación deberían crearse objetivamente como algoritmos, IA o aprendizaje profundo, por poner algunos términos de moda? Hasta cierto punto, un debate de este tipo parece llevarnos al borde del sinsentido, sin que nadie lo nombre explícitamente: "El emperador no lleva ropa".

Como todas las decisiones científicas, la elección de una prior debe justificarse en principio, ya sea según convenciones "objetivas" o "subjetivas" como un grado de convicción personal, no tiene mayor interés. Eso sería objeto de un debate crítico a la vista de un objeto de investigación concreto. Tanto si se utiliza una prior no informativo como una prior altamente informativo, ambos tienen consecuencias, tanto desde el punto de vista del contenido como desde el punto de vista matemático. Si la justificación matemática (y sea la de la Entropía Máxima y la teoría de la información) es siempre la mejor solución en términos de contenido y adaptada al caso concreto y al contexto, debe decidirlo de nuevo el discurso científico. Consideramos muy peligroso un procedimiento estándar automático que se aplique siempre en todas las condiciones. Quizá algo de esto pueda funcionar muy bien en las ciencias naturales. No podemos decir nada al respecto. En las ciencias sociales, donde prácticamente no hay leyes y sólo unos pocos planteamientos merecen realmente el título de "teoría", pero tienen como objeto de estudio entidades muy complejas, parece apropiado elegir una prior que pueda justificarse en términos de contenido. Lo que se discute raramente es que el cambio de la prior puede dar lugar a una pregunta diferente y, en consecuencia, a respuestas diferentes.

El criterio que nos guía es la transparencia, la reproducibilidad y la capacidad de mantener un discurso intersubjetivo sobre las decisiones y las consecuencias. Por último, pero no menos importante, esto se complementa con otras exigencias: divulgación de datos para reanálisis, estudios de replicación, etc. (Vasishth, Mertzen, Jäger & Gelman, 2018). También nos gustaría ver un poco de pensamiento dialéctico, de modo que las contradicciones no parezcan dignas de eliminación, sino que den lugar a una investigación más profunda para realizar la integración a un nivel superior, lo que suele significar un aumento de la complejidad.

El enfoque objetivo frente al subjetivo de Bayes no es en realidad una contradicción para nosotros, sino sólo una expresión de situaciones de información diferentes. Si no se dispone de información, parece justificada y razonable la elección según el Principio de Máxima Entropía (Jaynes, 1957a, 1957b) u otro enfoque que no fije un objetivo concreto. Sin embargo, si se dispone de información cualitativa, debe utilizarse e influir de forma plausible en la prior. Se trata de una característica, no de un error. Si una prior representa la convicción subjetiva del investigador, debe reflejarse y tenerse en cuenta, pero no debe ser un criterio de exclusión. Tampoco lo es el hecho de que no sea en absoluto trivial traducir el conocimiento cualitativo en una prior, es decir, una distribución matemática, de la forma más inequívoca posible.

La adopción o no de un punto de vista conservador (palabra clave: regularización, véase el capítulo 6.12) depende de la decisión de los investigadores a lo largo del caso individual concreto. Desde el punto de vista de la teoría científica, se podría argumentar: Si no se arriesga, no se gana. El conocimiento debe asumir el riesgo de la falsificación. Si la elección de la hipótesis previa es arriesgada, este intento puede fracasar, y eso es bastante destacable. Puede fracasar tanto más rápidamente si es simplemente el producto de una convicción subjetiva o de un capricho. El empirismo debería entonces tener un efecto devastador y eso está muy bien. Pero, a la inversa, el conocimiento subjetivo puede transformarse en conocimiento objetivo si demuestra ser científico y pertinente para la acción. No otra cosa persigue el programa de investigación "*Teorías subjetivas*" según Groeben y Scheele (1977), sin por ello perder una palabra sobre Bayes ni haberlo intentado nunca. Por desgracia, el propio Groeben (1986) rechaza vehementemente la reconstrucción de casos sensu Oevermann. Tal procedimiento no se aleja del procedimiento científico. Más bien, la inter-subjetividad más que objetividad – se logra cuando la aplicación de esta información cualitativa es transparente y está justificada, al igual que cualquier otro paso del proceso de investigación. Y debe ser revisable en el futuro.

Desde la perspectiva de las ciencias sociales, algunas partes de este debate deben distinguirse de las ciencias naturales. Por ejemplo, el objeto de la física, por muy matemáticamente complejos que sean los cálculos a nivel atómico e incluso a niveles más pequeños, es un objeto simple – comparado con la complejidad de un ser humano individual o incluso de un grupo o de unidades sociales aún mayores – (palabra clave: interpretación de Copenhague de la teoría cuántica). Así pues, si la física es el modelo estrella en cuanto a poder predictivo y precisión, es porque en última instancia la física se ocupa de hechos muy simples. Las personas y los colectivos son mucho más complejos, aunque ciertamente pueden describirse bien con modelos matemáticos dentro de ciertos límites. Pero por su propia naturaleza, a medida que aumenta la complejidad, la objetividad – operada como intersubjetividad justificada y discursiva basada en

datos – disminuye, no aumenta. Así pues, la física tiene un límite superior de precisión y objetividad al que las ciencias sociales no se acercan (no pueden acercarse). En las ciencias sociales, es mucho más necesario el tratamiento exacto de la imprecisión y la incertidumbre, que los métodos cualitativos pueden lograr en parte (véase el capítulo 11.2 sobre el análisis de secuencias o el capítulo 12 sobre la minimización booleana). La estadística de Bayes también ofrece esta posibilidad de tratar la incertidumbre.

En el caso de la aplicación de la estadística de Bayes, nos parece importante aceptar que el conocimiento previo es legítimo y se utiliza, es decir, todo lo que se sabe sobre un campo. La alternativa sería que los investigadores se pusieran artificialmente en un estado comatoso de ingenuidad y fingieran que el trabajo previo de ellos mismos y de otros investigadores es completamente irrelevante para todas las decisiones metodológicas y los cálculos numéricos. La subjetividad no es un problema si – como ya se ha mencionado – la elección de una distribución a priori se deriva del tema y está lógicamente vinculada y justificada. La investigación cualitativa está familiarizada con un problema comparable, por ejemplo en la elección de los códigos apropiados dentro del paradigma de codificación (véase el capítulo 9). Las comparaciones a nivel de codificación probablemente conducirían a un escaso acuerdo entre los investigadores. Sin embargo, la agregación hacia conclusiones debería ser muy similar o la misma entre distintos investigadores que trabajen de formas diferentes. No en vano, en este caso se abandonó el criterio de objetividad en favor de la *transparencia*, la *justificabilidad*, la *intersubjetividad*, la *reproducibilidad*, el *acceso a los datos originales*, etc. Afortunadamente, McShane, Gal, Gelman, Robert y Tackett (2019) no piden otra cosa "nueva" en su artículo, ahora también para la estadística. La transparencia no es suficiente, sin embargo, como señala Gelman (2017a).

6.6 El vínculo entre el teorema de Bayes y la estadística clásica

Un vínculo clave entre el teorema de Bayes y la estadística clásica es que esta última se ocupa exclusivamente de la *Likelihood* y, en general, solo de su máximo (máxima Likelihood según Ronald A. Fisher; Aldrich, 1997; Stigler, 2007), es decir, de un proceso de análisis puramente basado en datos. Sin embargo, según el teorema de Bayes, se examina un estado de la información basado en el contexto, es decir, de forma *multiperspectiva*. Los análisis no tienen lugar en condiciones infinitas (preferiblemente asintóticas), sino que siempre se refieren a condiciones finitas, independientemente de que la muestra sea grande o pequeña. Si la prior se trunca en una distribución uniforme, las estimaciones de la estadística clásica y bayesiana convergen o son asintóticamente idénticas si el teorema de Bernstein-von Mises es válido (Bickel & Kleijn, 2012). Entonces son principalmente los datos (= Likelihood) los que determinan el aspecto de la distribución posterior y la influencia de la prior tiende cada vez más a cero.

El teorema de Sergej Natanovic Bernštejn (1880-1968) y Richard von Mises (1883-1953) afirma que, para los modelos paramétricos, la distribución posterior en condiciones normales se concentra asintóticamente alrededor del parámetro verdadero independientemente de la distribución a priori. Así pues, la distribución posterior es aproximadamente normal, con valor medio y matriz de covarianza invertible (= matriz de información de Fisher). La matriz de información de Fisher, que se remonta a Ronald A. Fisher, proporciona información sobre la calidad óptima de las estimaciones de los parámetros y desempeña un papel importante en la estimación de máxima Likelihood. Debido a la naturaleza asintótica del teorema de Bernstein-von-Mises, la estadística bayesiana y la frecuentista están vinculadas. Esto significa que para los modelos paramétricos ambos enfoques, si están óptimamente alineados, conducen asintóticamente a los mismos resultados. Las excepciones y limitaciones del teorema son formuladas por Freedman (1963, 1965) y Diaconis y Freedman (1986), respectivamente, es decir, bajo qué condiciones el teorema de Bayes no conduce asintóticamente a la independencia de la prior.

6.6.1 Comprensión intuitiva de la probabilidad

Antes de examinar la intuición con más detalle, volvamos al principio de la estadística, es decir, al concepto de *probabilidad*. Esto distingue fundamentalmente la estadística clásica de la bayesiana. La estadística frecuentista se basa en el concepto de recuento y en el cociente de, por ejemplo, casos positivos en relación con la totalidad de todos los casos, es decir, una probabilidad relativa que se equipara a una probabilidad *objetiva*. El estudio de la estadística clásica demuestra que se practica una comprensión no intuitiva de la probabilidad (véase el capítulo 4.2.1). El intervalo de confianza clásico no proporciona información sobre la probabilidad con la que un parámetro se encuentre en un determinado intervalo de valores. Un valor *p* no dice nada sobre lo probable que es una hipótesis de interés a la luz de los datos empíricos, etc. Pero éstas son exactamente las afirmaciones que los usuarios esperan de la estadística y cómo se entienden generalmente las probabilidades en relación con la estadística. Para explorar la comprensión general de la estadística, basta con realizar una simple encuesta entre colegas y estudiantes sobre cómo se definen exactamente el valor *p* y el intervalo de confianza clásico según Neyman-Pearson (Gigerenzer, Krauss & Vitouch, 2004; Haller & Krauss, 2002). Por lo tanto, la estadística clásica suele aplicarse mal porque la comprensión de los parámetros es contraintuitiva.

En cambio, la estadística bayesiana proporciona una comprensión intuitiva. En el teorema de Bayes, la relación entre (*H*)ypótesis (o modelo) y (*D*)atos puede expresarse fácilmente.

$$p(H \mid D) = \frac{p(D \mid H) \cdot p(H)}{p(D)} \tag{6.38}$$

El teorema de Bayes nos permite evaluar la *probabilidad de las hipótesis* y, por tanto, comparar diferentes hipótesis, aunque Gelman y Shalizi (2010, 2013) afirman que no se trata de encontrar la mejor teoría/hipótesis, sino de integrar todos los enfoques sensatos entre sí en un modelo complejo e incluir casos especiales. La estadística clásica por otro lado se concentra en estimar p(D|H) y ni siquiera puede calcular p(H|D). Con Bayes, la *probabilidad* es ahora un potencial o una suposición sobre la ocurrencia de uno o varios sucesos incondicionales y esta probabilidad puede tener causas diferentes y combinables. Como se ha visto, la justificación de tal suposición puede ser el conocimiento experto disponible, pero también la intuición y la experiencia o simplemente datos empíricos. Algunos llaman a esto concepto de probabilidad *subjetiva*. El término subjetividad parece una mala elección para un juicio justificable porque, debido a las asociaciones e interpretaciones del término subjetividad – un individuo, un sujeto, posiblemente con emociones cambiantes, no accesibles emociones, visión interior subjetiva inaccesible, etc. – pasa por alto la ciencia. Todas estas subjetividades reales no son, por definición, ni ostensibles ni verdaderamente reconstruibles existentes. Parece mucho más sensato denotar esta variante de la probabilidad gris como un constructo multidimensional que se basa únicamente en una *argumentación fundada en el objeto* y que puede tener diversas causas. Aunque *la intuición* y la experiencia pueden lograr buenos resultados en la práctica, sólo se ajustan de forma limitada a esta definición de razonamiento. La intuición representa un arte superior, aparte del intelecto, que se basa en la experiencia interior y en un razonamiento directo (de esencia) y permite una visión directa de las cosas, pero sin que ésta sea argumentable o justificable. Por el contrario, una visión directa de las cosas ya no requiere el intelecto, sino más, sin entrar aquí en más detalles. *La experiencia*, a su vez, se basa en hechos pasados y sólo se ajusta a la definición si puede ser comunicada de forma argumentalmente condensada. Desde un punto de vista científico, la argumentación, es decir, el encadenamiento lógico de argumentos con tesis, antítesis, etc., es uno de los *puntos centrales* de cualquier planteamiento científico. Pensando en Kekulé y el descubrimiento del anillo de benceno en 1865 (Hussy, 1986) se puede afirmar que la intuición no sólo puede desempeñar un papel en la ciencia, sino es una de las capacidades más destacadas de los científicos. Pero cuando se trata de probabilidades, es esencial responder a la pregunta "¿Por qué se espera tal o cual suposición? y la intuición no puede hacerlo sin una lógica de justificación, aunque sea correcta en la práctica. Corresponde al individuo transformar intuición subjetiva en una cadena de argumentación intersubjetivamente verificable que se pueda discutir científicamente. Si este paso tiene éxito, si la demanda de argumentos se cumple, la intuición puede hacer una

contribución sustancial al tratamiento de las probabilidades. Sin embargo, en cuanto se dispone de una cadena de argumentación comunicable, y éste es un requisito mínimo para cualquier trabajo científico serio, ya no se puede hablar de subjetividad. El siguiente término de probabilidad utilizado en el contexto de los cálculos bayesianos por tanto nunca es subjetivo, porque no se limita al sujeto, ya que es intersubjetivamente verificable. Aunque difiere cualitativamente del de la estadística clásica, no se debe a una posible inaccesibilidad empírica. El problema real de la estadística clásica con esta probabilidad es que no se basa en el pasado finito contable, sino que representa una cantidad exible que puede llenarse de contenido en cualquier momento. Nos parece que la flexibilidad, que exige una gran responsabilidad y seriedad, es a veces demasiado para la gente y por eso la gente tiende a preferir una virtud relativa contable. No parece haber ningún contraargumento consistente por el que las suposiciones sobre la ocurrencia de acontecimientos deben basarse exclusivamente en frecuencias relativas pasadas y no también en estimaciones argumentativamente accesibles. En este punto, se puede citar a Hume, quien cuestionó críticamente que la experiencia del pasado diga algo sobre el futuro, ya que el pasado y el futuro no son directamente accesibles a los sentidos. Esto equipara la relativa frecuencia en términos de importancia a las valoraciones expertas y las cadenas de razonamiento.

Una nota importante: la probabilidad no debe confundirse con la causalidad. Un potencial es un espacio de posibilidad, no un espacio de realidad. El concepto de probabilidad es, por tanto, absoluto, ya que abarca todo el espacio de probabilidad, que se denota como el intervalo entre cero y uno. La causalidad, por su parte, se refiere a una teoría dirigida y qua teoría de las relaciones causa-efecto a lo largo del eje temporal, es decir, lo temporal anterior causa algo que es posterior en el tiempo. C.G. Jung (1875-1961) propuso el concepto de sincronicidad que relaciona la proximidad temporal de acontecimientos sin una relación causal directa, pero con un significado subjetivo. No profundizamos en este concepto.

La probabilidad puede, por ejemplo, indicar el grado de certeza de que existe una determinada causalidad. Pero no la condiciona ni interviene de forma efectiva. Se limita a describir las interrelaciones a nivel numérico respecto a la ocurrencia de dichas relaciones causa-efecto.

Otro término de probabilidad intuitivamente accesible se utiliza en física cuántica. En mecánica cuántica, la función de onda de una partícula describe su ubicación en relación con un área espacial. Y esto corresponde a la probabilidad de la ubicación, es decir, si la partícula acabará en esa zona o no. Esto no es sólo estadísticamente sino también *no determinista*. La función de onda describe, pues, la distribución de la probabilidad de la partícula de encontrarse en una zona del espacio. Si imaginamos el área espacial en paralelo y al mismo tiempo con sus diferentes probabilidades de ocurrencia, se obtiene una sensación de la probabilidad y del hecho de que no requiere cantidades contables. Al contrario, es sólo una pequeña parte. Podríamos preguntarnos en cualquier momento "¿Cuándo vamos a morir?" y la probabilidad se siente subjetivamente cada vez más alta a medida que envejecemos. Pero como sólo nosotros podemos morir cuando preguntamos por nosotros mismos, y eso ostensiblemente aún no ha tenido lugar en el pasado contablemente finito, no existe una relativa sabiduría relativa. Por supuesto, podríamos consultar una estadística de población para ver cuánto tiempo (hombre, mujer, dependiendo de la edad, clase social, nutrición, condición física, etc.) vivirá normalmente. Pero estos valores no tienen ningún efecto causal sobre nosotros, podríamos ser siempre una excepción. Así que podríamos morir de un cáncer no diagnosticado en quince días, de un accidente de coche esa misma tarde, o dentro de décadas, simplemente de un fallo cardíaco. En concreto no sabemos exactamente cuándo ocurrirá esto. Por lo tanto, ¿la probabilidad no es predecible o sólo debido a la población a la que pertenecemos? No, claro que no. Sería posible crear un modelo que, además de los factores mencionados en las estadísticas de población, incluyera información de nuestra biografía familiar o de nuestro estilo y forma de vida, que no son de naturaleza contable, pero que marcan nuestro estado de conocimiento previo. Con un poco de esfuerzo, se podría derivar de ello una estimación posterior que incluya tolerancias de incertidumbre hacia arriba y hacia abajo. Esto podría actualizarse con nuevos acontecimientos y cambios. Un día quedaría entonces claro si es más bien la estadística global de la población o el razonamiento ajustado individualmente podría predecir nuestro tiempo real de muerte. Queda por ver quién es mejor y más preciso en sus predicciones. De lo que se trata simplemente es mostrar que no sería sensato ignorar la información disponible simplemente porque no representa una precisión relativa. Estructuralmente, tanto las estadísticas de población como la previsión cualitativa individual son legítimas. Con Bayes, las estadísticas de población se toman como otra fuente legítima y podrían combinarse con los valores individuales para formar una posterior común.

Una probabilidad tiene una distribución de probabilidad, es decir, qué valores en qué rango son más o menos probables. Una distribución de probabilidad posterior en general proporciona información sobre la distribución de probabilidad con la que se estima un parámetro. Con Bayes, los parámetros se tratan como *variables aleatorias con una distribución de probabilidad asociada*, mientras que en la estadística clásica se consideran los parámetros como valores fijos de modo que las imprecisiones de medición y de otro tipo no hacen más que enmascarar sus *verdaderos valores*. El punto de máxima densidad de esta distribución posterior equivale al punto de máxima probabilidad (= MAP, es decir, máxima a posteriori), es decir, el *mejor estimador de un parámetro dados los datos*, es decir, $p(H|D)$ si H = hipótesis (sobre el parámetro) y D = datos empíricos. La incertidumbre se caracteriza por el hecho de que existen límites de tolerancia y no sólo el punto de máxima probabilidad. Establecemos estos límites de forma similar a los de una prueba de significación idealmente sobre la base de consideraciones de contenido (véase el concepto ROPE según Kruschke, 2017a, capítulo 6.8.4.2). En la práctica, como es lógico, prevalece la convención, por ejemplo, un intervalo de confianza del 95 % que es cualitativamente diferente para Bayes que para Neyman-Pearson (véase el cap. 4.3.3.6). El intervalo de confianza bayesiano de un parámetro se denomina *intervalo creíble* (= IC) o intervalo de máxima densidad (Highest Density Interval = IDH) y designa con precisión el intervalo de valores en el que se encuentra un parámetro con la probabilidad seleccionada (Kruschke, 2012b, 2015b). El IDH es una especialización del IC, ya que todos los valores dentro del IDH tienen una probabilidad posterior mayor que cualquier valor fuera del IDH. Por supuesto habría que justificar la amplitud de un espectro de probabilidad adecuado (por ejemplo, el 95 %) sin acabar en convenciones incuestionables – y reabriendo así la discusión que ya tuvimos con la estadística clásica.

Un pequeño ejemplo muestra lo que está en juego. Generamos datos a partir de una distribución normal posterior. Con precisión τ conocida, la distribución normal es su propia distribución prior conjugada (Wikipedia, 2019e; Schnäbele, 2011). Como priores establecemos valores comunes como por ejemplo del rango CI con $\mu_{prior} = 100$ y $\tau_{prior} = 1/10^2$. Para los datos empíricos generados aleatoriamente con un tamaño de muestra $N = 100$ se aplica lo siguiente: $x = 108$ y $s = 9$ (`ptII_quan_Bayes_posterior.r`).

```
# https://en.wikipedia.org/wiki/Conjugate_prior
# normal with known precision tau
# POST:
# mean was estimated from observations with total precision
# (= sum of all individual precisions)
# t 0 {\displaystyle \tau _{0}} \tau _{0} and with
# sample mean μ 0 {\displaystyle \mu _{0}} \mu _{0}
set.seed(2745)
samp.n <- 100
samp <- rnorm(n=samp.n, mean=108, sd=15)
reps <- 4
ndraws <- 1e+4
mu.prior <- 100
s2.prior <- 10^2
tau.prior <- 1/s2.prior
tau.samp <- 1/var(samp)
mu.post <- ( tau.prior*mu.prior + tau.samp*sum(samp) ) /
(tau.prior + samp.n*tau.samp)
tau.post <- tau.prior + samp.n*tau.samp
s.post <- sqrt(1/tau.post)
theta.post <- matrix(rep(rnorm(ndraws, mu.post, s.post),
                  each=reps), ncol=reps,
                  nrow=ndraws, byrow=FALSE)
str(theta.post)
head(theta.post)
# repeat with same specs, but without same seed
theta.post2 <- matrix(rep(rnorm(ndraws, mu.post, s.post),
                  each=reps), ncol=reps,
                  nrow=ndraws, byrow=FALSE)
head(theta.post2)
```

Cuatro cadenas MCMC están disponibles aquí para la distribución posterior, un extracto:

```
> head(theta.post2)
    [,1] [,2] [,3] [,4]
[1,] 109 109 104 106
[2,] 109 109 104 106
[3,] 109 109 104 106
[4,] 109 109 104 106
[5,] 106 109 107 105
[6,] 106 109 107 105
```

Los valores priores μ_{prior} y τ_{prior} así como los datos empíricos dan los valores posteriores μ_{post} y τ_{post}

```
> mu.post
[1] 107
> tau.post
[1] 0.367
> s.post
[1] 1.65
```

Estos valores sirven de base para todas las demás extracciones aleatorias de la distribución posterior. Estas extracciones pueden resumirse y examinarse más de cerca, por ejemplo con los paquetes de R `coda` o `bayesplot`.

```
> summary(as.mcmc(theta.post))
Iterations = 1:10000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
plus standard error of the mean:
     Mean SD   Naive SE Time-series SE
[1,] 107 1.63  0.0163   0.0304
[2,] 107 1.66  0.0166   0.0295
[3,] 107 1.66  0.0166   0.0313
[4,] 107 1.64  0.0164   0.0354

2. Quantiles for each variable:
     2.5% 25% 50% 75% 97.5%
var1 104  106 107 108 110
var2 104  106 107 108 110
var3 104  106 107 108 111
var4 104  106 107 108 111
```

Para la salida gráfica (véase la Fig. 6.6) es adecuada la función `plotPost()` del paquete R `BEST`. Además de un histograma, permite la visualización del intervalo de confianza HDI (s. cap. 6.8.4.1) con envergadura `credMass`, un valor de comparación `compVal` así como el `ROPE` (s. cap. 6.8.4.2), que define una zona de equivalencia aproximada, es decir, un rango dentro del cual, por ejemplo, las diferencias se consideran muy improbables. A continuación, se examina si los valores posteriores cubren este rango o no, o en qué medida (Kruschke, 2011b, 2018). El `ROPE` se corresponde aproximadamente con las pruebas de equivalencia clásicas (Kruschke, 2016a, 2017a; Lakens, 2017c), pero es significativamente más flexible y espera una derivación basada en el contenido del objeto de investigación.

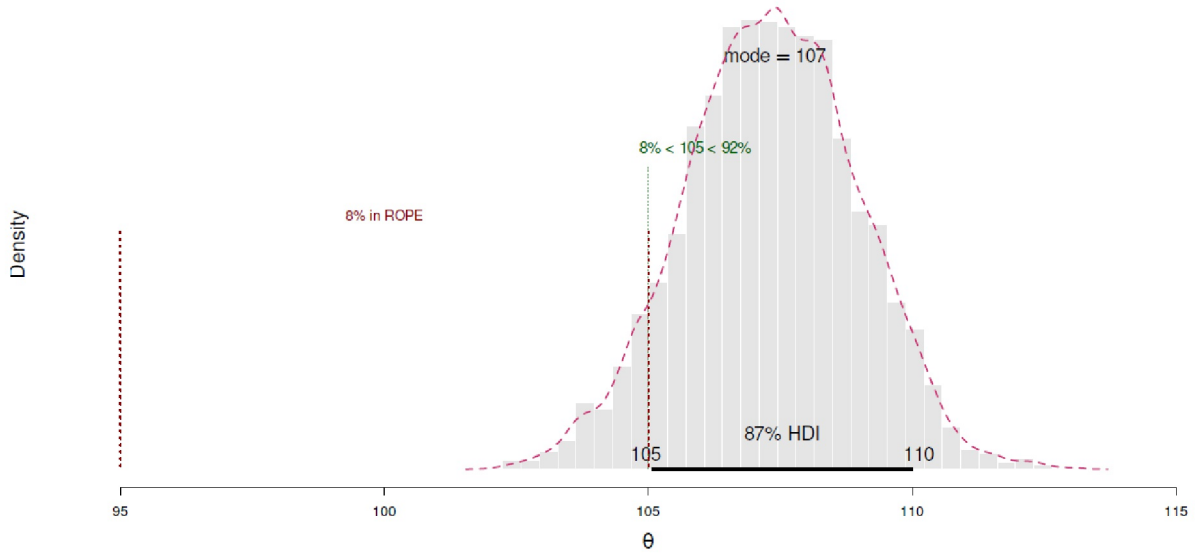


Figura 6.6: *Distribución posteriora*

```
plotPost(theta.post[,1], xlim=c(95,115), credMass=0.87,
        compVal=105, ROPE=c(95,105), showMode=TRUE,
        col="grey90", border="white",
        xlab=expression(theta), ylab="Density")
lines(density(theta.post[,1]), col="violetred3", lwd=2, lty=2)
```

Para seguir el curso de las extracciones, utilizamos `plot()` de `coda`. Otros gráficos bayesianos de interés de coda son `gelman.plot()`, `geweke.plot()`, `autocor.plot()` o `cumuplot()`, todos los cuales no tienen mayor interés aquí, ya que la posterior es una distribución normal para la que se conocen todos los parámetros del presente ejemplo (véase la Fig. 6.6). Las fórmulas para calcular los parámetros posteriores μ_{post} y τ_{post} son de conocimiento común (Taboga, 2010) y, por tanto, los gráficos no muestran sorpresas. Lo mismo puede decirse de diagnósticos más avanzados como `geweke.diag()`, `raftery.diag()`, `heidel.diag()` y `gelman.diag()`, respectivamente, que se tratan con más detalle en el capítulo 6.13.3. Estas pistas de diagnóstico resultan ser relevantes cuando se trata de la estimación MCMC con una distribución posterior desconocida. Todos ellos proporcionan información diferente sobre la cuestión si los resultados MCMC indican una distribución estacionaria, es decir, si las cadenas convergen o si las cadenas están muy correlacionadas entre sí a lo largo del tiempo. Por ejemplo, el diagnóstico Gelman-Rubin (Gelman & Rubin, 1992; Brooks y Gelman, 1998) estima la diferencia de varianza *entre* las cadenas MCMC frente a *dentro* de las cadenas MCMC relacionadas con cada parámetro del modelo. Si aquí hay una discrepancia mayor, se asume la no convergencia. El diagnóstico MCMC es un campo muy amplio. Básicamente, se trata de determinar si las muestras son aleatorias por naturaleza (en términos de sus dependencias a lo largo del tiempo) y si los parámetros estimados son independientes entre sí. Diferentes cadenas deberían conducir a los mismos resultados y estimaciones de parámetros. Las señales de alerta son básicamente correlaciones y autocorrelaciones existentes, así como un aspecto inusual de los posteriores o incluso diferencias sistemáticas entre cadenas y dentro de cadenas individuales. Así pues, observamos las cadenas MCMC (véase la Fig. 6.7):

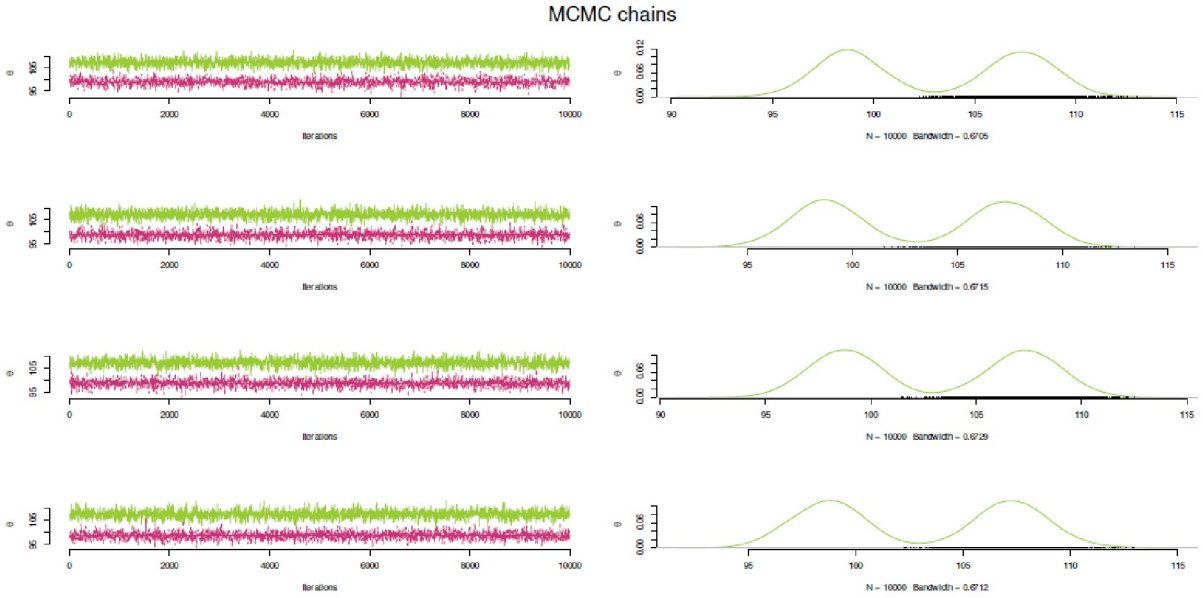


Figura 6.7: *Cadenas MCMC (diagnóstico)*

```
# combine two MCMC chains lists, because gelman.plot()
# expects this
mcmc.l <- mcmc.list(as.mcmc(theta.post),as.mcmc(theta.post2))
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l")
plot(mcmc.l, type="l", col=c("yellowgreen","violetred3"), bty="n",
                                  smooth=TRUE, ylab=expression(theta))
mtext("MCMC chains", outer=TRUE, line=-1, cex=1.5, side=3)
```

Son posibles otros gráficos, de los cuales sólo imprimimos el de Gelman-Rubin-Brooks en la Figura 6.8:

```
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l")
# Gelman plot
gelman.plot(mcmc.l)
mtext("Gelman plot", outer=TRUE, line=-2, cex=1.5, side=3)
# autocorrelation plot
acf(theta.post)
# cumulative quantile plot
par(mfrow=c(2,2))
cumuplot(theta.post, type="l", col="darkred", bty="n",
           ylab=expression(theta),auto.layout=FALSE)
```

Y se puede acceder a una selección de otros diagnósticos (en parte con gráficos, no impresos) a través de:

```
# Geweke's convergence diagnostic R-Code
geweke.diag(theta.post)
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l")
geweke.plot(as.mcmc(theta.post))
mtext("Geweke plot", outer=TRUE, line=-1, cex=1.5, side=3)

# Raftery and Lewis's diagnostic
raftery.diag(theta.post, q=0.025, r=0.005, s=0.95)

# Heidelberg and Welch's diagnostic
heidel.diag(theta.post)
autocorr.diag(as.mcmc(theta.post))
```

```
crosscorr(mcmc.l)
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l")
crosscorr.plot(mcmc.l)
mtext("Cross-Correlation plot", outer=TRUE, line=-1, cex=1.5, side=3)
effectiveSize(mcmc.l)
# Gelman and Rubin's convergence diagnostic gelman.diag(mcmc.l,
# multivariate=FALSE) gelman.diag(mcmc.l, multivariate=TRUE)
# MCMC chain plots
xyplot(mcmc.l[,1:4])
acfplot(mcmc.l[,1:4], lag.max=200)
lapply(mcmc.l, cor)
```

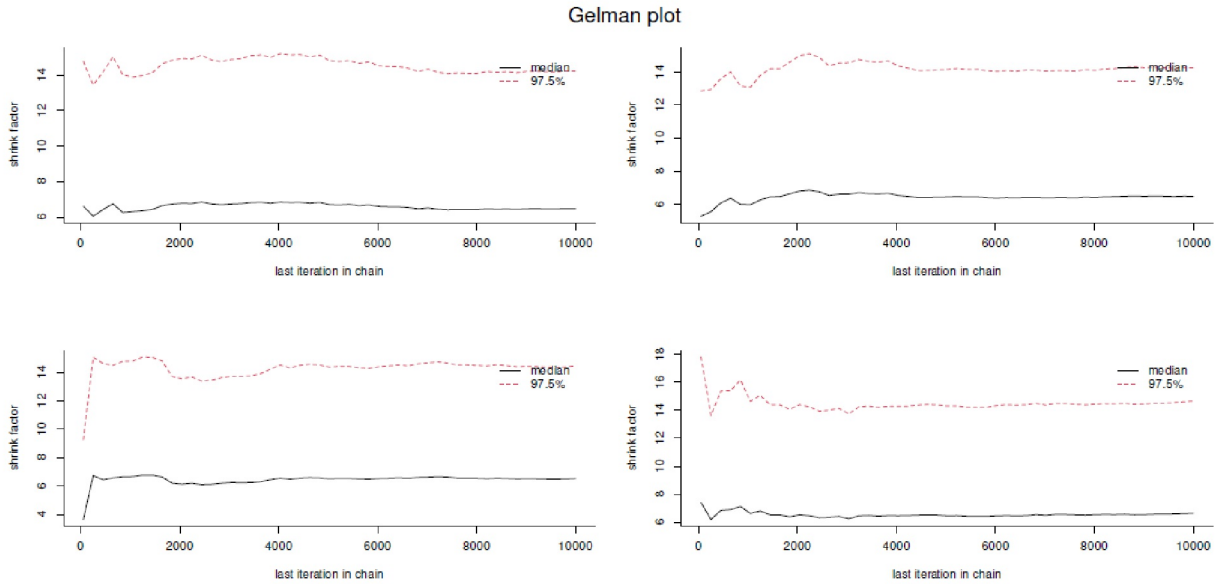


Figura 6.8: *Gráfico Gelman-Rubin-Brooks (Diagnóstico MCMC)*

El objetivo de la mayoría de los gráficos y criterios de diagnóstico es comprobar si las cadenas MCMC convergen o no con el tiempo y evaluar la estabilidad de la variable posterior resultante. Encontrará más detalles y bibliografía en las páginas de ayuda de las funciones R correspondientes. Sin embargo, lo que se puede afirmar con la conciencia tranquila es que el IDH contiene el parámetro θ buscado con una probabilidad de `credMass`, que en este caso es el valor medio, pero puede ser cualquier cantidad de interés. La estadística clásica no puede hacer esta afirmación. La salida de `hdi()` de `HDInterval` da los valores inferior y superior del intervalo por sorteo. La anchura del intervalo puede cambiarse dependiente del interés, por ejemplo, del típico 95% al 87% o 67% o 99% u otros valores límite que esperemos estén justificados en términos de contenido. Los intervalos de confianza bayesianos (s. cap. 6.8.4.1) resultan por cadena MCMC en

```
> hdi(theta.post, credMass=0.87)
       [,1] [,2] [,3] [,4]
lower   105  105  105  105
upper   110  110  110  110
attr(,"credMass")
[1] 0.87
```

Lo importante de estos argumentos es darse cuenta de que las estimaciones de probabilidad son en realidad siempre afirmaciones de rangos (de valores) y que detrás de cada coeficiente o estimación puntual hay un rango que le pertenece pero que es menos probable. Por eso no se mira solo al MAP (= máximo a posteriori, véase McElreath, 2015), sino al aspecto de toda la posterioridad. En nuestra opinión, la estadística

debería en principio manejar todas las estimaciones como declaraciones sobre rangos de valores, incluso o especialmente cuando son estimaciones puntuales. Dado que operamos con muestras finitas y nunca encontramos condiciones asintóticas de infinito en la realidad o distribuciones teóricamente conocidas tales que los errores estándar sean cero, toda estimación tiene algo de imprecisión. Sin embargo, la imprecisión puede tratarse con precisión, como muestran las combinaciones lógicas de la lógica difusa / Fuzzy Logic (véase el capítulo 12.7). Trabajar con rangos de valores también requiere considerar, tanto en el rango límite superior como en el inferior, hasta qué punto los valores son plausibles y cómo se relacionan con la respuesta a la pregunta de investigación. La ventaja desde nuestro punto de vista es que esta comprensión nos acerca a la realidad, que cambia constantemente y cuyas relaciones causa-efecto multicomplejas suelen ser difíciles de captar más allá de los estudios de laboratorio.

Por último, un ejemplo de datos abstractos. Generamos datos distribuidos normalmente y especificamos una prior. El paquete `Bolstad` de R ofrece una función `normgcp()`, que puede utilizarse para obtener tanto la distribución Posterior como la Likelihood. Las trazamos con nuestra propia función llamada `plot.mean.post()`. Como parámetros de los datos aleatorios elegimos N = 10, $\mu = 6.5$ y $\sigma = 2$. Como prior elegimos una distribución normal – justificada en este caso – y el valor x = 6.5 para la media y s = 2 para la desviación estándar, es decir, los valores con los que se generaron los datos. Conocemos la distribución (`ptII_quan_Bayes_simple-estimation-mean-post.r`).

```
# posterior distribution of a mean
seed <- 3856
set.seed(seed)
N <- 10
MW <- 6.5
SD <- 2
x <- rnorm(N, MW, SD)
# informed prior
PRIOR <- c(6.5,2)
bp.m.res <- normgcp(x, sigma.x=sd(x), density="normal",params=PRIOR, plot=FALSE)#.alt
str(bp.m.res)
#plot.mean.post(bp.m.res)
```

y una estimación media de la Posterior de

```
> # output
> mean(bp.m.res)
[1] 6.67
```

Como se puede ver en la Figura 6.9 (izquierda), los máximos de las distribuciones respectivas (a priori, Likelihood, posterior) coinciden con bastante exactitud. Esto no es sorprendente, ya que hemos especificado una probabilidad a priori que se corresponde con los datos, de modo que la probabilidad a priori y la probabilidad a posteriori no difieren realmente. En la práctica, esto no ocurriría. Dado que los valores teóricos de μ y σ son conocidos, pueden entrar en la prior como informativos. En otras palabras, tenemos hipótesis razonadas muy sólidas sobre μ y σ. Modifiquemos esto. Supongamos teóricamente que no sabemos exactamente cuál es la verdadera distribución y supongamos para la prior $\mu = 4$ y $\sigma = 3$. Los datos empíricos permanecen inalterados, sólo cambiamos nuestras suposiciones. La llamada a la función va precedida de una llamada a `layout()` y se pasa la opción `add=TRUE` para comparar los dos gráficos. Por último, el eje X se escala al mismo rango de valores.

```
# changed informed prior
PRIOR1 <- c(4,3)
bp.m.res1 <- normgcp(x, sigma.x=sd(x), density="normal", params=PRIOR1, plot=FALSE)
mean(bp.m.res1)
str(bp.m.res1)
layout(matrix(1:2,ncol=2), width = c(2,2),height = c(1,1))
plot.mean.post(bp.m.res, xlim=c(0,14), add=TRUE)
plot.mean.post(bp.m.res1, xlim=c(0,14), add=TRUE)
# comparison
bp.m.res1$likelihood == bp.m.res$likelihood
```

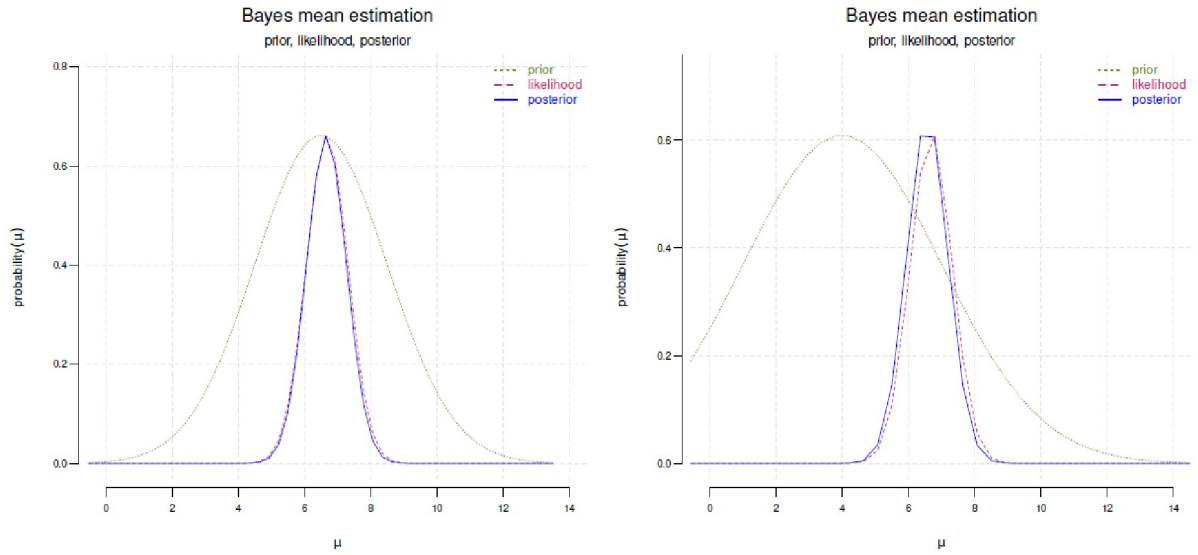


Figura 6.9: *Media según estimación Bayes (Prior, Likelihood, Posterior)*

Ahora la salida correspondiente de los valores medios de los dos posteriors con varios decimales (véase la Fig. 6.9, derecha), así como la comparación resultante

```
> # output
> mean(bp.m.res)
[1] 6.66646312
> mean(bp.m.res1)
[1] 6.56902606
> mean(bp.m.res)/mean(bp.m.res1)
[1] 1.0148328
```

Ahora destacan varias cosas. En primer lugar, la prior se ha desplazado claramente hacia la izquierda, lo que no es sorprendente, también es mucho más amplia, lo que también corresponde al cambio en la desviación estándar. Por otra parte, la Likelihood es idéntica, ya que los datos siguen siendo los mismos. La posterior ha cambiado ligeramente. Los valores medios de las variables posteriores sólo son ligeramente diferentes a pesar de los diferentes priores, por lo que la posterior actúa como un compromiso entre la Prior y Likelihood.

¿Qué aprendemos de esto? A pesar de un claro cambio en la prior (menor media, mayor desviación estándar), la Likelihood determina en gran parte la Posterior. No obstante, la probabilidad a priori "corrige" o "ajusta" la probabilidad a posteriori. También se podría decir que las hipótesis a priori y los datos empíricos, que por muchas razones como las hipótesis a priori pueden ser suposiciones, se compensan entre sí para producir una estimación plausible teniendo en cuenta la información disponible en el momento t. La información de μ y σ para la Prior es una *Prior informativa*, que en principio son datos pasados, estimaciones de expertos u otras fuentes fiables. Los detalles son poco importantes, en el caso de los estudios reales, por supuesto. Pasemos ahora a una *Prior no informativa*, por ejemplo, una *distribución uniforme* da la misma probabilidad inicial a cada valor del intervalo de valores. Eso significa que a priori hay una distribución equitativa, de modo que ningún valor o rango de valores se ve favorecido. Es decir, todo es posible, nada se excluye ni se prefiere. Y eso es una expresión de ignorancia, pero también de suposiciones erróneas, si, por ejemplo, se consideran posibles valores que pueden ser (casi) excluidos o son imposibles.

```
> # non-informative prior
> bp.m.res2 <- normgcp(x, sigma.x=sd(x), density="uniform",
+              params=NULL, plot=FALSE)
Known standard deviation :2.002
```

```
> mean(bp.m.res)
[1] 6.666463
> mean(bp.m.res2)
[1] 6.68314
> mean(bp.m.res)/mean(bp.m.res2)
[1] 0.9975046
> layout(matrix(1:2,ncol=2), width = c(2,2),height = c(1,1))
> plot.mean.post(bp.m.res2)
```

No ha cambiado mucho el gráfico (véase la Fig. 6.10, sin escalar, es decir, sin ajustar la altura de las distribuciones entre sí) ni las estimaciones. Lo mismo ocurre con la comparación de los valores medios. Una prior poco o nada informada puede ser mejor que una prior informada pero formulada de forma totalmente errónea. Por otra parte, una Prior bien informada y bien fundamentada es siempre preferible, aunque conduzca a una mayor incertidumbre Posterior. Esta última es entonces una realidad justificada y no representa un problema metodológico en sentido estricto. Si los datos empíricos en sí mismos sólo tienen un bajo contenido de información, sólo suelen dar lugar a pocas conclusiones inequívocas.

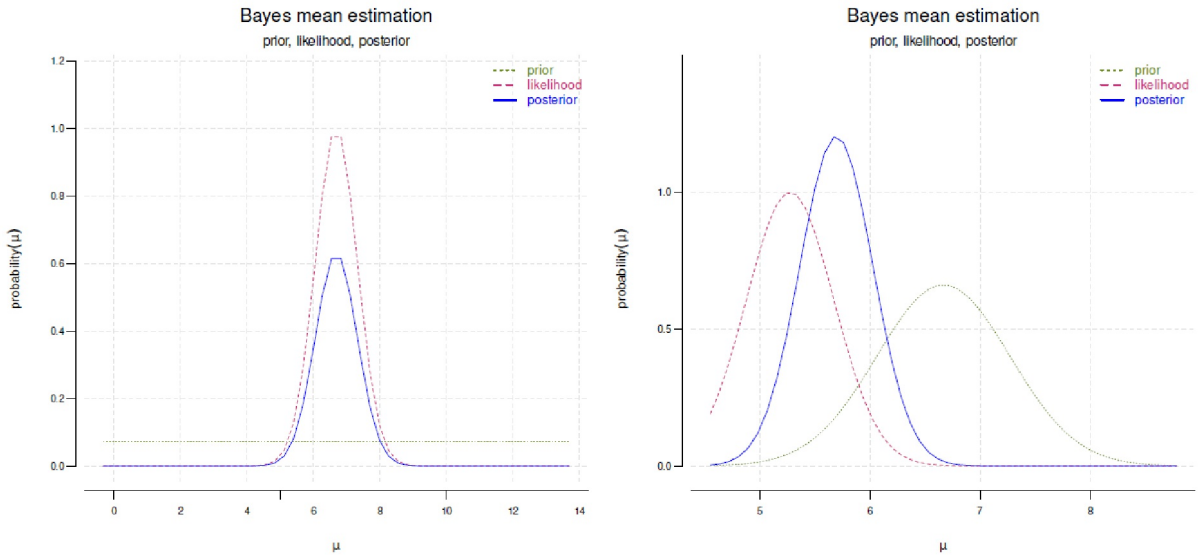


Figura 6.9: *Media según estimación Bayes (Prior no informada, Likelihood, Posterior, sin escalado)*

Como puede observarse, la variable Posterior no permanece constante, sino que está sujeta a fluctuaciones en función de la variable Prior. Sin embargo, cuantos más datos haya disponibles, menos suele influir la Prior. Por supuesto, esto puede cambiar en función del modelo, el tamaño de la muestra, el tipo de datos y los métodos utilizados. No siempre cabe esperar cambios tan pequeños como en este caso. En primer lugar es comprender que la Prior, la Likelihood y la Posterior son tres distribuciones distintas que difieren entre sí, la Prior y la Likelihood influyendo a su vez en la Posterior. Ahora podríamos repetir esto y tomar las estimaciones de la primera Posterior para usarla como Prior para la estimación basada en datos nuevos y aleatorios con los mismos parámetros.

```
> # replication
> # new data
> # prior = posterior of round 1 (=bp.m.res)
> PRIOR3 <- c(mean(bp.m.res), sd(bp.m.res))
> x2 <- rnorm(N, MW, SD)
> bp.m.res3 <- normgcp(x2, sigma.x=sd(x2), density="normal",
+              params=PRIOR3, plot=FALSE)
Known standard deviation :2.333
> mean(bp.m.res3)
[1] 6.568651
> mean(bp.m.res)
[1] 6.666463
> mean(bp.m.res3)
```

```
[1] 6.568651
> mean(bp.m.res)/mean(bp.m.res3)
[1] 1.014891
> plot.mean.post(bp.m.res3,add=TRUE)
```

Ahora resulta obvio que los nuevos datos cambian bastante las distribuciones. La Likelihood representa los nuevos datos aleatorios y la Posterior es una mezcla de Likelihood (= nuevos datos) y Prior (= anterior Posterior, véase más arriba). La Posterior se sitúa entre la Likelihood y la prior. Ahora bien, este proceso podría repetirse durante el tiempo que se desee y, por ejemplo, la diferencia en los valores medios de la posterior por ejecución o acumulativamente a lo largo de todas las ejecuciones podría almacenarse y representarse gráficamente. Esto podría dar una impresión de cómo la estimación de la distribución de la Posterior de las Posteriores evoluciona hacia una estimación robusta a pesar del muestreo aleatorio constante – en aproximación a los parámetros poblacionales que conocemos. Los lectores interesados pueden implementar esto como una tarea con R.

Dado que la Prior actual corresponde a la Posterior inicial anterior, es idéntica a ella. Utilizar la Posterior como una Prior para los mismos datos no tendría sentido y distorsionaría masivamente los resultados y llevaría a una inconsistencia en el enfoque. Por lo tanto, debe evitarse. También hay que tener en cuenta que la elección de una Prior reduce la gama de valores. Si, por ejemplo, se establece la Prior = cero en determinados puntos o se excluye de un rango de valores, este rango no se tendrá en cuenta en la Posterior, ya que la Posterior es un producto de Prior * Likelihood y si una parte del producto es cero, la otra parte ya no cambia nada sobre el cero. Desde el punto de vista del Prior, no hay masa en este rango de valores. Por tanto, la reducción del intervalo de valores debe hacerse con mucho cuidado y sólo cuando se esté seguro. Por ejemplo, se podría cortar una estimación de la esperanza de vida humana por debajo de cero, porque los años negativos no se producen; y a > 130 años o un poco menos, porque ningún ser humano ha llegado nunca a esa edad. Estas expectativas serían legítimas en el marco de los conocimientos biológicos sobre el cuerpo humano.

En el presente ejemplo, la variable Prior pertenece a *la misma familia de distribuciones* que la variable Posterior. Ambas tienen una distribución normal. Esto no tiene por qué ser así, y en la práctica los tipos de distribución de la Prior y la Posterior difieren considerablemente en función del modelo de análisis utilizado y los parámetros del modelo asociados, las variables implicadas, etc. En el caso de que el tipo de distribución de la Priori sea igual a la Posterior, se habla de *densidad a priori conjugada a la Likelihood*. Esto es matemáticamente muy elegante porque no hay ningún cambio. A nivel práctico, esto puede significar que, en primer lugar, las propiedades de la distribución son conocidas y el teorema de Bayes puede resolverse idealmente de forma analítica en lugar de mediante simulación, ya que sólo se actualizan los parámetros de la distribución y se omite el laborioso cálculo del denominador del teorema de Bayes (= probabilidad total). En sentido estricto, sin embargo, no ocurrirá en la práctica que la distribución de una Prior se conozca con exactitud. Pero en muchos casos una determinada distribución puede ser una muy buena aproximación de la misma y la pérdida de información asociada debida a la diferencia entre aproximación y realidad es soportable. En la práctica, probablemente no exista una única Prior "verdadera" conocida y se trabaje exclusivamente con aproximaciones a la realidad. Todo es un modelo.

6.7 Posibilidades de los métodos de análisis de datos

Las posibilidades de los procedimientos de análisis de datos no difieren con Bayes de las de la estadística frecuentista. En principio, se dispone del mismo arsenal de análisis. Sin embargo, las afirmaciones e interpretaciones resultantes difieren considerablemente – entre otras cosas, en el concepto de probabilidad utilizado, la(s) hipótesis estadística(s) investigada(s) y, en consecuencia, en toda la argumentación cualitativa, así como en la forma en que pueden adquirirse sucesivamente los conocimientos. Mientras que la estadística frecuentista define y asigna la probabilidad a través de la frecuencia relativa de los sucesos, las muestras

deben considerarse aleatorias e independientes entre sí, en la estadística bayesiana todo recibe una distribución de probabilidad. Esto incluye datos, parámetros y, en consecuencia, también modelos e hipótesis. Las muestras pueden basarse unas en otras, lo que corresponde al aprendizaje a partir de la experiencia. Además, la estadística bayesiana se ocupa de la probabilidad de las hipótesis, mientras que la estadística clásica se ocupa de la probabilidad de los datos ante una hipótesis nula. Los parámetros no tienen distribución en la estadística clásica, sino se les consideran fijos y sólo están *oscurecidos* por el error de medición y otras magnitudes (por ejemplo, error clásico y teoría de la medición de la psicología, Lienert & Raatz, 1998). Las hipótesis tampoco tienen probabilidad de ser estimadas, ya que el valor *p* se refiere a la probabilidad de los datos o datos más extremos dada la hipótesis (nula) y no al revés (véase el capítulo 4.3.9.1, Tschirk, 2014). Ambos están permitidos explícitamente por la estadística de Bayes. La discusión sobre el concepto de probabilidad es aún más amplia y está relacionada con la *subjetividad* (véase el capítulo 6.3.1) y *deducción frente a inducción* (véanse los capítulos 2.2 y 2.1). Varios autores desarrollan estas discusiones con mucho más detalle, por ejemplo Jaynes (2003).

Observamos que la estadística bayesiana puede realizar cualquier análisis estadístico concebible al modo bayesiano según el teorema de Bayes. En la forma de llegar a las conclusiones entran en juego diversos métodos de trabajo (como la estimación por intervalos o la comprobación de hipótesis). Éstos van acompañados de diversos criterios de calidad que no se corresponden con los de la estadística clásica. Pasamos ahora al ámbito de los métodos de trabajo.

6.8 Visión general de los métodos de trabajo bayesianos

A continuación, examinamos la forma en que se procesan y abordan los problemas en la estadística bayesiana y la base sobre la que esto tiene lugar. Existe una división casi natural de los métodos de trabajo en las áreas de

- Factores de Bayes y comprobación de hipótesis de Bayes (véase el capítulo 6.8.1)
- criterios de información (s. cap. 6.8.2)
- modelización y estimación por intervalos (véase el capítulo 6.8.4)

que trataremos con más detalle a continuación. Dependiendo del método de trabajo, existen áreas más detalladas como

- funciones de pérdida (véase el cap. 6.8.1.2)
- el sobreajuste y el infraajuste (véase el capítulo 6.8.3),
- calidad predictiva de los modelos (comprobaciones predictivas posteriores y evaluaciones gráficas, véase el capítulo 6.8.4.3).

Además, se presentan ejemplos de investigación en forma de digresiones.

La estadística de Bayes no trata de valores *p*, significancias y pruebas de hipótesis (nulas), α- y β- tasas de error, etc. El *núcleo de los análisis* es la estimación y explicación de los parámetros en el marco de su distribución de probabilidad (Jaynes, 2003) en forma de inferencia consistente (Studer, 1996b). Así, a cada parámetro de un modelo y a cada cada hipótesis investigada se les puede asignar una distribución de probabilidad con un punto de máxima densidad que represente la mejor estimación dada la información *Prior* * *Likelihood* disponible. A partir de ahí, se pueden determinar probabilidades para los parámetros, los modelos y las hipótesis, lo que también permite contrastar los modelos entre sí – si es necesario, utilizando los denominados factores de Bayes (Ly, Verhagen & Wagenmakers, 2016), que se remontan a Jeffreys (1939/1961). El estadístico Andrew Gelman critica esto en muchos artículos, afirmando que no se trata de contrastar hipótesis entre sí. Más bien, debería centrarse en el ajuste de modelos para ampliarlos de modo

que también puedan explicar casos especiales. Por tanto, *estimar en lugar de probar* (pero véase Wagenmakers, Lee, Rouder & Morey, 2019 sobre las diferencias especiales respecto a la comparación/prueba de modelos y la estimación de parámetros, similar a la conocida paradoja de Lindley-Bartlett, véase la sección 4.4.14.2). Los modelos resultantes se examinan cada vez más gráficamente y con la ayuda de compro-baciones predictivas posteriores (Kruschke, 2013c; Gabry, Simpson, Vehtari, Betancourt & Gelman, 2019). Las estimaciones de intervalos complementan la visión de los modelos. Las comprobaciones de modelos surgen como simulaciones de la distribución posterior. Estas simulaciones se comparan con los datos empíricos disponibles en términos de desviaciones que indican si el modelo no se ajusta y dónde. Entonces comienza la revisión del modelo. Mediante este proceso iterativo, los modelos se ajustan para explicar los datos existentes y predecir conjuntos de datos nuevos o simulados. Este enfoque se aleja de la simple comprobación de hipótesis punto por punto y se acerca a la modelización compleja a través de variables y parámetros. El enfoque de estimar en lugar de probar ya es conocido de la estadística frecuentista (Gelman & Hill, 2007). La ventaja es que se utiliza más información y la significación es más amplia, pero con más incertidumbre. Por ejemplo, una prueba de hipótesis que termina con *aceptado frente a rechazado* puede parecer más clara a primera vista. Pero también contiene bastantes incertidumbres, que, sin embargo, se transforman en pseudo-certidumbres mediante la aplicación de barreras críticas (palabra clave: pruebas de significación).

De este modo, las hipótesis (modelos) pueden contrastarse entre sí, lo que equivale a la clásica prueba de significación. La pregunta de investigación que requiere precisamente una aplicación de este tipo debe ser la que conduzca al objetivo. El punto crítico es que las pruebas siempre implican barreras que distinguen entre *no significativo y significativo*, ya sea de forma continua o en gradaciones cualitativas. Y estas barreras deben fijarse adecuadamente para el caso y justificarse en función del objeto de investigación. Este es un enfoque que no se puede esperar fácilmente y siempre ni de Bayes ni de los estadísticos clásicos (McShane, Gal, Gelman, Robert & Tackett, 2019 y Tramow et al., 2018 respectivamente en respuesta al artículo de Benjamin et al., 2018-01-01). Incluso con Bayes existe el peligro de aplicar normas descontextualizadas debido al deseo de un procedimiento universal de análisis automático de datos libre de contexto (Gigerenzer & Marewski, 2015). Además del ajuste del modelo y los factores de Bayes, se dispone de más información específica como criterios de calidad en el contexto de la estadística de Bayes para evaluar modelos y compararlos entre sí. Ahora también se utilizan en la estadística frecuentista, por ejemplo, para evitar los valores *p*.

6.8.1 Factores de Bayes y prueba de hipótesis de Bayes

> „Andrew Gelman wishes to state that he hates Bayes factors. The reasons for his aversion are detailed in Gelman & Rubin [40] and Gelman et al. [41, ch. 6] and mainly concern the practice of assigning prior mass to a single point from a continuous distribution, and the resulting sensitivity to the prior distribution (see also [42]).“
>
> (Andrew Gelman en Marsman, Schönbrodt, Morey, Yao, Gelman & Wagenmakers, 2017, p.5, nota 3).

No se puede acusar a Andrew Gelman de ocultar su postura sobre los factores de Bayes. Los factores de Bayes se remontan a Jeffreys (1935, 1939/1961). Mediante los factores de Bayes (BF), las hipótesis H_i o modelos M_i pueden contrastarse entre sí para decidir qué modelo puede explicar mejor los datos (Kass & Raftery, 1993; Ly, Verhagen & Wagenmakers, 2016). Entonces, si se dan *D*(atos) y dos hipótesis de interés H_1 y H_2 (o denominadas modelos M_1 y M_2), el teorema de Bayes se aplica a H_1 y H_2 respectivamente.

$$p(H_1|D) = \frac{p(D|H_1) \cdot p(H_1)}{p(D)} \tag{6.39}$$

$$p(H_2|D) = \frac{p(D|H_2) \cdot p(H_2)}{p(D)} \tag{6.40}$$

Ambas ecuaciones pueden resolverse para *p(D)* e igualarse. La probabilidad total se omite entonces. Una transformación da como resultado

$$\frac{p(H_2|D)}{p(H_1|D)} = \frac{p(D|H_2) \cdot p(H_2)}{p(D|H_1) \cdot p(H_1)} \tag{6.41}$$

Esto corresponde a

$$\text{Posterior odds} = BF_{21} \cdot \text{Prior odds} \tag{6.42}$$

o resuelto según el factor de Bayes BF_{21}

$$\text{BF}_{21} = \frac{p(D|H_2)}{p(D|H_1)} \tag{6.43}$$

$$= \frac{p(H_2|D) \cdot p(H_1)}{p(H_1|D) \cdot p(H_2)} \tag{6.44}$$

$$= \frac{p(H_2|D) \cdot p(H_1)}{p(H_2) \cdot p(H_1|D)} \tag{6.45}$$

y para mayor claridad

$$= \left[\frac{p(H_2|D)}{p(H_2)}\right] \cdot \left[\frac{p(H_1)}{p(H_1|D)}\right] \tag{6.46}$$

Tabla 6.6: *Factores de Bayes y su significado según Jeffreys (1939/1961, p.432)*

$\log_{10}(B_{10})$ Factor de Bayes	Factor de Bayes B_{10}	Evidencia frente H_0
< 0	< 1	negativo, pro H_0
0 ... 0.5	1 ... 3.2	no vale mencionarle
0.5 ... 1	3.2 ... 10	sustancial
1 ... 1.5	10 ... 32	fuerte
1.5 ... 2	32 ... 100	muy fuerte
> 2	> 100	decisivo

Estos tres términos, *Prior Odds, factor de Bayes* y *Posterior Odds*, tienen significados diferentes.

Prior Odds – este término representa el conocimiento previo y cuantifica en forma de ratio la Likelihood relativa de H_2 frente a H_1 *antes* de que se observaran los datos *D*. La Prior tiene en cuenta los conocimientos anteriores procedentes de estudios previos, búsquedas bibliográficas, conocimientos de expertos, etc. Al parecer, el propio Harold Jeffreys prefería la hipótesis de que ambos modelos (hipótesis) son igualmente probables y entonces asignó $p(H_2) = p(H_1) = 0.5$. Esto significa que el 50% de la masa a priori se asigna a un único punto – aquí cero, es decir, ninguna diferencia entre H_2 frente a H_1 – y el resto se distribuye en un rango válido de valores (postulato de simplicidad / "simplicity postulate" de Wrinch-Jeffreys, Wrinch & Jeffreys, 1921, 1923).

Posterior Odds – este término corresponde a las *Prior Odds después* de se han filtrado los datos *D*, es decir, una vez que se han multiplicado las *Priori Odds* por el *factor de Bayes* (= parte de Likelihood). Representa la relación entre la plausibilidad relativa de las dos hipótesis contrapuestas H_2 frente a H_1 dados los datos.

BF21 – es el *factor de Bayes* $p(D|H_2)=p(D|H_1)$ y representa la relación de las probabilidades relativas de los datos observados *D* bajo cada una de las dos hipótesis H_2 y H_1 respectivamente. Un factor de Bayes de *BF21 = 5*, por ejemplo, significa que los datos tendrían 5 veces más probabilidades de estabilizarse bajo la hipótesis H_2 que bajo la hipótesis H_1. La escala del factor de Bayes continuo fue dividida en clases cualitativas por Jeffreys (1939/1961, Apéndice B, p.432). Jeffreys propuso una división a lo largo de medias unidades de la escala log10. Un factor de Bayes inferior a uno, por ejemplo BF21 = 0.2, significa lo contrario. Aquí los datos serían 5 veces más probables bajo la hipótesis H_1 en lugar de H_2. El propio Jeffreys (ibíd., Apéndice B) clasificó la importancia de los *BF* de tal forma que los valores BF21 > 10 son pruebas sólidas para una hipótesis H_2 frente a H_1, mientras que BF21 > 3 son, en el mejor de los casos, buenos para "informes andecdóticos" o no merece la pena informar. Los valores intermedios (ibíd., p. 256) se consideran interesantes para los jugadores, pero no lo bastante significativos para un artículo científico. Como señalan otros autores en este punto (Ly, Verhagen & Wagenmakers, 2016, p.10) los estudios estadísticos clásicos con valores p > 0.01 *no* suelen alcanzar este límite.

Tabla 6.7: *Factores de Bayes y su importancia según Kass y Raftery (1995, p.777).*

$2 \cdot \log_e(B_{10})$ Factor de Bayes	Factor de Bayes B_{10}	Evidencia frente H_0
< 0	< 1	negativo, pro H_0
0 ... 2	1 ... 3	no vale más que mencionarlo
2 ... 6	3 ... 20	positivo
6 ... 10	20 ... 150	fuerte
> 10	> 150	muy fuerte

Un $BF_{21} = 1$ conduce a una situación claramente indiferente que requiere más datos o exige un cambio en los modelos. La interpretación de estos valores no es fácilmente trasladable a las ciencias sociales. Por lo tanto, es necesario aclarar muy detalladamente qué significado tienen los valores en la realidad de la psicología, las ciencias de la educación, sociología, etc. Lo mismo ocurre con las Lo mismo cabe decir de las interpretaciones abstractas de los tamaños del efecto propuestas por Cohen (1992), que son heurísticamente interesantes pero no deben tomarse al pie de la letra en todos los contextos. Otra ayuda a la interpretación convierte el valor BF_{21} en decibelios, ya que este valor es fácilmente comprensible para los humanos. Según Good (1979), en la vida cotidiana las personas perciben cambios de 1 decibelio o un bit de 1/3, lo que corresponde a un odds ratio de 5/4. Como alternativa a este esquema existe una clasificación simplificada según Kass y Raftery (1995). Las tablas 6.6 y 6.7 recogen respectivamente los valores sugeridos por Jeffreys (1939/1961, Apéndice B, p.432) y Kass y Raftery (1995, p.777).

Se pueden utilizar algunos ejemplos de datos para ilustrar el trabajo con los factores de Bayes. En primer lugar, recapitulemos la paradoja de Lindley-Bartlett (Lindley, 1957; Bartlett, 1957, véase el capítulo 4.4.14.2).

Siguiendo una recomendación de Defazio (2016-09-13), esta se puede relativamente fácil de reproducir tomando una Prior vago, un tamaño de muestra grande y un valor medio que sea $2/\sqrt{n}$ y una desviación estándar de 1 (`ptII_quan_Bayes_BayesFactors_test-hypos.r`).

```
# Lindleys Paradox
set.seed(66755)
n <- 5e+4
daten <- rnorm(n, mean=2/sqrt(n), sd=1)
```

Los valores generados aleatoriamente corresponden a las expectativas (gráfico no impreso).

```
hist(daten, prob=T, pre.plot=grid())
lines(density(daten), col="darkred")
summary(daten)
sd(daten)
```

El tamaño del efecto, es decir *d de Cohen* es muy pequeño.

```
> # Cohen's d
> (mean(daten)-0)/1
[1] 0.0102885
```

Si ahora realizamos la prueba *t* clásica para comprobar la hipótesis nula

```
> # classical t-test
> t.test(daten)
One Sample t-test
data: daten
t = 2.3137, df = 49999, p-value = 0.02069
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
0.001572603 0.019004395
sample estimates:
mean of x
0.0102885
```

y lo compara ción `ttestBF()` del paquete `BayesFactor` de R, la diferencia se hace evidente. Elegimos una hipótesis nula no especificada para BF_1 y una prueba unilateral para BF_2.

```
> # unspecific null hypothesis
> bf1 <- ttestBF(daten)
> # one-sided test delta < 0 versus point null
> bf2 <- ttestBF(daten, nullInterval=c(-Inf,0))

> bf1
Bayes factor analysis
--------------
[1] Alt., r=0.707 : 0.07331898 ±0.02%
Against denominator:
Null, mu = 0
---
Bayes factor type: BFoneSample, JZS

> bf2
Bayes factor analysis
--------------
[1] Alt., r=0.707 -Inf<d<0 : 0.001517911 ±0%
[2] Alt., r=0.707 !(-Inf<d<0) : 0.1451201 ±0%
Against denominator:
Null, mu = 0
---
Bayes factor type: BFoneSample, JZS
```

Dentro de un análisis, podemos comparar las respectivas hipótesis para formar un propio factor de Bayes:

```
> # delta < 0 versus delta > 0
> bf2[2]/bf2[1]
Bayes factor analysis
--------------
[1] Alt., r=0.707 !(-Inf<d<0) : 95.60512 ±0%
Against denominator:
Alternative, r = 0.707106781186548, mu =/= 0 -Inf<d<0
Bayes factor type: BFoneSample, JZS
```

Ahora seleccionamos el intervalo contra el que aún no hemos realizado la prueba:

```
> # same from opposite direction
> bf3 <- ttestBF(daten, nullInterval=c(0,Inf))
> bf3
Bayes factor analysis
--------------
[1] Alt., r=0.707 0<d<Inf : 0.1451201 ±0%
[2] Alt., r=0.707 !(0<d<Inf) : 0.001517911 ±0%
Against denominator:
Null, mu = 0
---
Bayes factor type: BFoneSample, JZS
> bf3[1]/bf3[2]
Bayes factor analysis
--------------
[1] Alt., r=0.707 0<d<Inf : 95.60512 ±0%
Against denominator:
Alternative, r = 0.707106781186548, mu =/= 0 !(0<d<Inf)
---
Bayes factor type: BFoneSample, JZS
```

Para no confundir la situación siempre desfavorable de una Prior vago y un efecto pequeño, se pueden crear dos pruebas unilaterales en lugar de una hipótesis nula bilateral. Una Prior vaga con efectos pequeños conduce a rangos de valores que reciben una probabilidad a priori que no puede justificarse a nivel de contenido. Esto conduce a pruebas conservadoras, es decir, un efecto pequeño necesita una muestra muy grande para ser considerado en absoluto. Esto aleja el problema de la paradoja de Lindley de los factores de Bayes en el análisis detallado y lo acerca a la elección de la Prior, que no puede elegirse de forma objetiva y completamente inespecífica. Si al planificar un experimento, por ejemplo, ya se presuponen efectos pequeños, entonces los efectos previstos deberían representarse con la mayor precisión posible en la Prior. Esperar un efecto pequeño, pero actuar como si pudiera ser cualquier cosa al elegir la Prior, es contraproducente y es más probable que sea una expresión de una planificación defectuosa del diseño y del trabajo teórico a priori. En el paquete `BayesFactor` de R la Prior se especifica en la mayoría de los casos como una Prior Cauchy con un parámetro de escala de r = 0.707 y es cuestionable si esto es siempre apropiado en términos de contenido (Schimmack, 2015b, 2015c) o lo que esto realmente tiene que ver con el contexto. Defazio (2016-09-13, p.5), señala,

> „The default Cauchy prior used in the BF t-test essentially says that you expect a 50 %chance that the absolute effect size (Cohen's *d*) is larger than 1, which is a completely implausible in many settings."

La elección de los factores de Bayes como instrumento de análisis por sí sola no basta, ni mucho menos, para cumplir los requisitos de un diseño (experimental) completo. Es necesario justificar que el instrumento de análisis seleccionado se corresponde plenamente con el diseño y sus condiciones específicas. Si un factor de Bayes compara dos hipótesis que parecen muy improbables a la vista de los datos, no se obtendrá nada significativo. El nivel del factor de Bayes no proporciona información al respecto, pero una evaluación cuidadosa requiere una visión amplia del diseño y su realización.

Para los dos factores de Bayes resultantes, se puede formar una relación que demuestra de forma impresionante que los intervalos a la derecha y a la izquierda de cero no reciben los mismos factores de Bayes. A continuación, las dos hipótesis BF_1 y BF_2 se contrastan entre sí en un nuevo modelo, y las posibilidades de prueba resultantes se presentan claramente en forma de tabla:

```
> bfall <- c(bf1,bf2)
> bfall
Bayes factor analysis
--------------
[1] Alt., r=0.707             : 0.07331898  ±0.02%
[2] Alt., r=0.707 -Inf<d<0    : 0.001517911 ±0%
[3] Alt., r=0.707 !(-Inf<d<0) : 0.1451201   ±0%
Against denominator:
Null, mu = 0
---
Bayes factor type: BFoneSample, JZS
> plot(bfall)
>
> bfmat <- bfall/bfall
> bfmat
                          denominator
numerator                 Alt., r=0.707             Alt., r=0.707 -Inf<d<0
Alt.,   r=0.707           1.00000000                48.30256
Alt.,   r=0.707 -Inf<d<0  0.02070284                 1.00000
Alt.,   r=0.707 !(-Inf<d<0) 1.97929729              95.60512
                         denominator
numerator                 Alt., r=0.707 !(-Inf<d<0)
Alt., r=0.707                           0.50522981
Alt., r=0.707 -Inf<d<0                  0.01045969
Alt., r=0.707 !(-Inf<d<0)               1.00000000
```

Defazio (2016-09-13, p.5) comenta este contexto, „The extreme case of this is when neither hypothesis includes the true value of θ. In such cases, the *BF* will generally not converge to any particular value, and may show strong evidence in either direction." No es de extrañar que Defazio describa sus argumentos siguiendo el estudio ya comentado de Bem (2011a) y Bem, Utts y Johnson (2011, véanse los capítulos 4.4.2.2 y 6.8.1.6) y las respuestas de Wagenmakers, Wetzels, Borsboo y van der Maas (2011). Al hacerlo, el autor (Defazio, 2016-09-13, p.5) critica tanto el estudio de Bem como los contraargumentos, como Wagenmakers, Wetzels, Borsboo y van derMaas (2011) y Wagenmakers, Wetzels, Borsboo, Kievit y van derMaas (2011) que colocan una Prior grande sobre un rango improbable de tamaños de efecto de tal manera que incluso los defensores de la clarividencia no supondrían efectos tan grandes. Por el contrario, los resultados de Bem muestran la naturaleza de los factores bayesianos, es decir, que una hipótesis extremadamente improbable – la clarividencia – es cambiada por los datos de Bem a una hipótesis (todavía muy) improbable. En términos absolutos, la clarividencia sigue siendo improbable precisamente porque fallaron las réplicas externas (Ritchie, Wiseman & French, 2012) y, por tanto, toda la base de datos sigue siendo escasa.

Los factores de Bayes *sólo* proporcionan información *sobre el cambio en las expectativas previas a través de los datos* y no dicen *nada* sobre el nivel global del modelo desde el punto de vista del Posterior. Esta es tarea de una aplicación bayesiana completa, que *no* incluye los factores de Bayes. Los BF "sustituyen" tan fácilmente a los valores *p* precisamente porque, como puede verse en las calibraciones de los valores *p* (Sellke, Bayarri & Berger, 2001, véase también el capítulo 6.8.1.4), los amplían mediante la elección de una Prior, pero básicamente intentan hacer lo mismo: pruebas puras comparando integrales. Si la integral del factor de Bayes se sustituye por las respectivas estimaciones MLE, se reduce a una prueba de ratio de Likelihood. Con Priores iguales, el factor de Bayes corresponde al cociente de las probabilidades posteriores de los modelos que se van a probar entre sí. Esto se relaciona con la cuestión de si el supuesto de igualdad de Priores está justificado o hasta qué punto los distintos modelos están justificados en función de distintos supuestos a priori.

La proximidad de los factores de Bayes al enfoque frecuentista puede ilustrarse con un pequeño ejemplo. Sirvan como datos las edades en el estudio de Gürtler (2005), que se examina con más detalle en el capítulo 6.8.4.6. En este estudio, se preguntó cualitativamente a adolescentes de secundaria, bachillerato y formación profesional acerca de sus experiencias y opiniones sobre el humor en el aula con un cuestionario que constaba de siete preguntas. Se analiza si estos grupos difieren en términos de edad. Un examen gráfico de los datos sugiere que es así (`ptII_quan_Bayes_BayesFactors_test-hypos.r`).

```
# read data from school study R-Code
diss <- read.table(file="LG_school-words-raw.tab",
sep="\t", header=TRUE)
dim(diss)
head(diss)
tail(diss)
diss.red <- subset(diss, stype %in% c("R","G"))
selvars <- c("ID","age","sex","stype")
diss.red <- diss.red[,selvars]
dim(diss.red)
naids <- which(is.na(diss.red), arr.ind=TRUE)
naids
diss.red.nona <- diss.red[-naids,]
age <- diss.red.nona$age
stype <- factor(diss.red.nona$stype)
sex <- diss.red.nona$sex
dats <- data.frame(age,sex,stype)
dats$age.jit <- age.jit <- jitter(age)
head(dats)
```

En primer lugar, nos encontramos con el problema de que las edades, si se dan en años, son muy poco específicas y olvidan que el tiempo es una variable continua. Si carecemos de la fecha de nacimiento, podemos crear artificialmente una variable pseudocontinua `age.jit` a partir de `age` que comparta los estadísticos de resumen de la edad. El caso individual es aquí menos importante, en la medida en que no sufrimos grandes pérdidas de conocimiento. Toleramos el pequeño aumento de la incertidumbre por la introducción de la pseudocontinuidad. Como puede verse, se mantiene dentro de los límites. La función de R `jitter()` es un medio de elección en este caso. Más adelante complementaremos esto con otra posibilidad para hacer los datos aún más continuos.

```
> # tables
> table(age)
age
14 15  16 17 18 19
34 120 94 20 40 9
> table(stype)
stype
G  R
46 271
> # not shown
> # table(age.jit)
> tapply(age,stype, summary)
$G
Min. 1st Qu. Median Mean 3rd Qu. Max.
17   18      18     18   18      19
$R
Min.  1st Qu. Median Mean  3rd Qu. Max.
14.00 15.00   15.00  15.44 16.00   19.00
> tapply(age.jit,stype, summary)
$G
Min.  1st Qu. Median Mean  3rd Qu. Max.
```

```
16.98 17.88   17.96  18.01 18.15   19.17
$R
Min.  1st Qu. Median Mean  3rd Qu. Max.
13.81 14.90   15.12  15.43 16.00   18.94
> tapply(age,stype, sd)
G R
0.5577734 0.9282000
> tapply(age.jit,stype, sd)
G         R
0.5620272 0.9387060
```

Los resultados se presentan gráficamente en forma de gráficos de caja (véase la Fig. 6.11):

```
# boxplots
TITLE <- "Study Gürtler"
SUB <- "age ~ sex"
par(mfrow=c(1,2))
bpx <- boxplot(age ~ stype, plot=FALSE)
bxp(bpx, notch=TRUE, ylab="age", xlab="age versus sex",
                          main="", frame=FALSE,
                          boxfill=c("green","darkorange"), border=2/2)
mtext(TITLE, 3, line=2.5, cex=1.5)
mtext(SUB, 3, line=1, cex=1.1)
bpx <- boxplot(age.jit ~ stype, plot=FALSE)
SUB <- "jitter(age) ~ sex"
bxp(bpx, notch=TRUE, ylab="age", xlab="age versus sex",
                                  main="", frame=FALSE,
                                  boxfill=c("green","darkorange"), border="darkblue")
mtext(TITLE, 3, line=2.5, cex=1.5)
mtext(SUB, 3, line=1, cex=1.1)
```

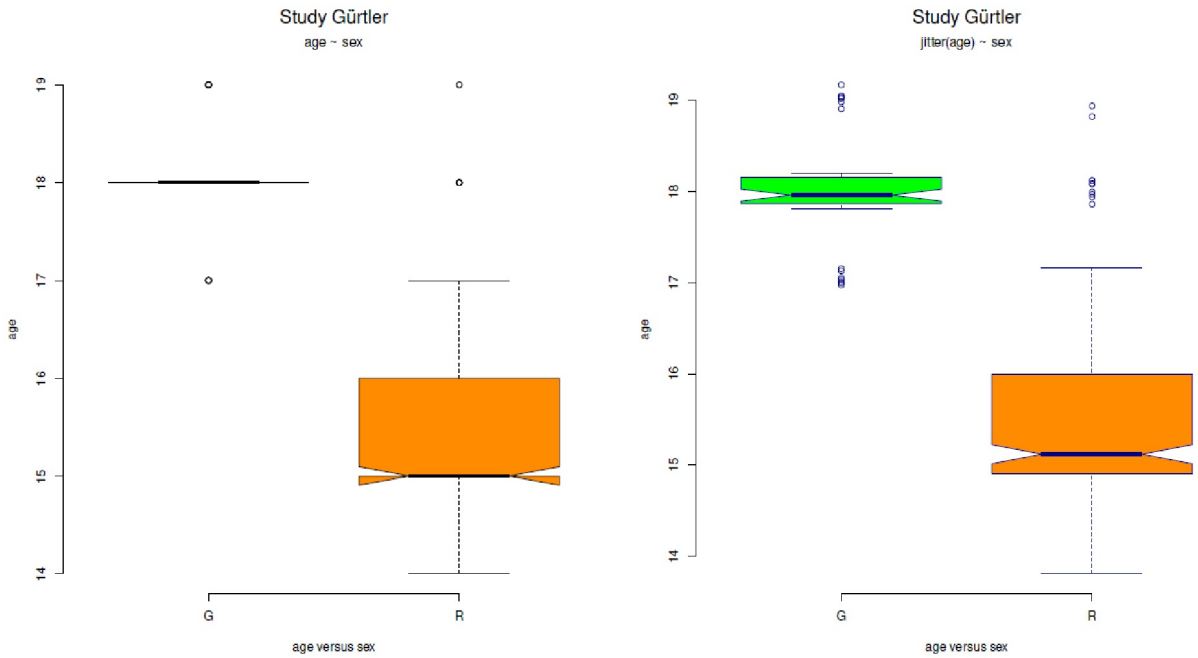


Figura 6.11. *Estudio de Gürtler (2005, boxplots edad frente a sexo, con y sin* `jitter`*).*

La prueba t frecuentista, la fuerza d del efecto y la prueba t con factor de Bayes indican algo similar. Nos limitamos a la variable pseudocontinua `age.jit` para la salida:

```
> # age.jit
> # simple t-test classical
> t.test(age.jit ~ stype, var.eq=FALSE)
Welch Two Sample t-test
data: age.jit by stype
t = 25.65, df = 94.186, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
2.380428 2.779866
sample estimates:
mean in group G   mean in group R
18.01002          15.42987
> # effect size
> cohensd(age.jit[stype == "G"], age[stype == "R"])
d|mean sd d|pooled sd
-3.355493 -2.908448
> # Bayes Factor t-test age.jit versus school type
> bf2s.1.jit <- ttestBF(formula=age.jit ~ stype, data=dats)
t is large; approximation invoked.
> bf2s.1.jit
Bayes factor analysis
--------------
[1] Alt., r=0.707 : 7.560753e+46 ±0%
Against denominator:
Null, mu1-mu2 = 0
---
Bayes factor type: BFindepSample, JZS
> bf2s.2.jit <- ttestBF(formula=age.jit ~ stype,
+ data=dats, nullInterval=c(-Inf,0))
t is large; approximation invoked.
t is large; approximation invoked.
t is large; approximation invoked.
t is large; approximation invoked.
> bf2s.2.jit
Bayes factor analysis
--------------
[1] Alt., r=0.707 -Inf<d<0    : 0.0009240192 ±NA%
[2] Alt., r=0.707 !(-Inf<d<0) : 1.512151e+47 ±NA%
Against denominator:
Null, mu1-mu2 = 0
---
Bayes factor type: BFindepSample, JZS
> bf2s.2.jit[1]/bf2s.2.jit[2]
Bayes factor analysis
--------------
[1] Alt., r=0.707 -Inf<d<0 : 6.11063e-51 ±NA%
Against denominator:
Alternative, r = 0.707106781186548, mu =/= 0 !(-Inf<d<0)
---
Bayes factor type: BFindepSample, JZS
> # check MCMC chains
> samps1.jit <- posterior(bf2s.1.jit, iterations=1e+5)
0%  10   20   30   40   50   60   70   80   90  100%
|----|----|----|----|----|----|----|----|----|----|
**************************************************|
> plot(samps1.jit, bty="n", col="violetred3")
Drücke Eingabetaste für den nächsten Plot:plot(bf2s.1.jit)
> summary(samps1.jit)
Iterations = 1:1e+05
Thinning interval = 1
Number of chains = 1
Sample size per chain = 1e+05
```

```
1. Empirical mean and standard deviation for each variable,
plus standard error of the mean:
             Mean       SD        Naive SE  Time-series SE
mu           16.7142    7.181e-02 0.0002271 0.0003966
beta (G - R)  2.5646    1.440e-01 0.0004555 0.0007974
sig2          0.8076    6.471e-02 0.0002046 0.0002072
delta         2.8607    1.974e-01 0.0006244 0.0009629
g            52.1665    1.576e+03 4.9831649 4.9831649
2. Quantiles for each variable:
             2.5%    25%     50%    75%      97.5%
mu           16.5730 16.6660 16.714 16.7624  16.8559
beta (G - R)  2.2824  2.4674 2.565   2.6614   2.8477
sig2          0.6911  0.7625 0.804   0.8489   0.9437
delta         2.4753  2.7272 2.860   2.9941   3.2480
g             1.1519  3.0981 6.234  14.9301 181.3894
```

Ahora podría comprobarse mediante factores de Bayes si tiene sentido incluir la variable sexo como factor distintivo en el modelo. Esto sigue la sugerencia de Kruschke (2013c) de que las pruebas del modelo deben ser bayesianas.

```
> dats$sex <- as.factor(dats$sex)
> str(dats)
'data.frame': 317 obs. of 4 variables:
$ age :    int 16 14 15 15 15 15 15 15 15 16 ...
$ sex :    Factor w/ 2 levels "m","w": 1 1 2 1 2 2 1 1 1 2 ...
$ stype :  Factor w/ 2 levels "G","R": 2 2 2 2 2 2 2 2 2 2 ...
$ age.jit:  num 15.9 14.1 15.1 15.1 14.8 ...
> # Bayes Factor t-test age versus sex
> ttestBF(formula=age.jit ~ sex, data=dats)
Bayes factor analysis
--------------
[1] Alt., r=0.707 : 0.3218293 ±0%
Against denominator:
Null, mu1-mu2 = 0
---
Bayes factor type: BFindepSample, JZS
```

Obviamente, esto no tiene sentido. Tampoco el clásico. En lugar de un factor, la pregunta podría ser ahora si "sexo" debe tratarse como un nivel separado (efecto aleatorio). Según la viñeta de `BayesFactor`, esto puede hacerse.

```
> # Bayes Factor anova age versus (school type + sex)
> anovaBF(formula=age.jit ~ stype + sex, data=dats)
|============================ 25%t is large; approximation invoked.
|============================================================| 100%
Bayes factor analysis
--------------
[1] sex                      : 0.3218293 ±0%
[2] stype                    : 7.560753e+46 ±0%
[3] sex + stype              : 1.164721e+46 ±2.05%
[4] sex + stype + sex:stype  : 2.605628e+45 ±2.05%
Against denominator:
Intercept only
---
Bayes factor type: BFlinearModel, JZS
> # with random effect = sex
> anovaBF(formula=age.jit ~ stype + sex, data=dats, whichRandom="sex")
|============================================================| 100%
Bayes factor analysis
--------------
```

```
[1] stype + sex : 3.584164e+46 ±1.91%
Against denominator:
age.jit ~ sex
---
Bayes factor type: BFlinearModel, JZS
```

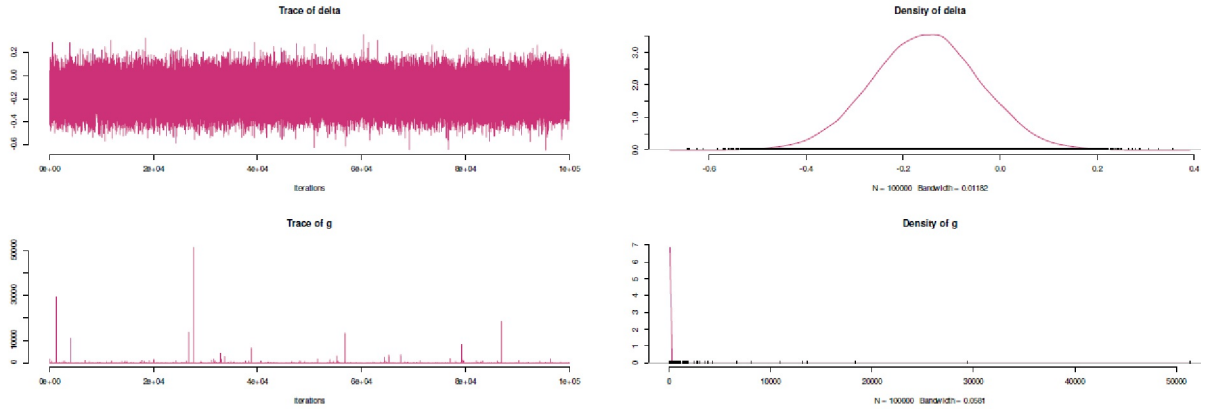


Figura 6.12. *Estudio de Gürtler (2005, muestras de la Posterior en el cálculo del factor de Bayes)*

Una vez más, no parece que esto prometa ninguna ganancia significativa de conocimiento. Pero ahora nos preguntamos si el modelo describe adecuadamente los datos. Eso no puede deducirse de los factores de Bayes. Así pues, a partir del modelo, que sólo contiene el tipo de escuela y ha resultó ser superior sobre la base de los factores de Bayes, se toman muestras de la Posterior, se representan gráficamente y se resumen (véase la Fig. 6.12).

```
# plot posterior samples R-Code
bf.best <- ttestBF(formula=age.jit ~ sex, data=dats)
bf.best.post <- posterior(bf.best, iterations=1e+5)
plot(bf.best.post, bty="n", col="violetred3")
```

Aquí no se observan problemas. ¿Es adecuado el modelo para predecir los datos? Comparemos esto con `brm()` del paquete `brms` de R. Empezamos con la variable `age`.

```
# check against Stan with brms R-Code
# full Bayesian
# not really good...
brm1 <- brm(age ~ stype + (stype|sex), data=dats)
summary(brm1)
plot(brm1)
pp_check(brm1, nsamples=100) # horrible with age as a category!
# not really good...
brm2 <- brm(age ~ stype + (1|sex), data=dats)
summary(brm2)
plot(brm2)
```

Ahora se realiza una comprobación predictiva posterior (posterior predictive check; véase cap. 6.8.4.3) con `pp_check()`:

```
pp_check(brm2, nsamples=100) # horrible with age as a category!
```

y esta demuestra que el modelo sólo describe inadecuadamente los datos. El resumen no parece satisfactorio en lo que respecta a la calidad del modelo de `brm1`:

```
> summary(brm1)
Family: gaussian
Links: mu = identity; sigma = identity
Formula: age ~ stype + (stype | sex)
Data: dats (Number of observations: 317)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
total post-warmup samples = 4000
Group-Level Effects:
~sex (Number of levels: 2)
                      Estimate  Est.Error   l-95% CI u-95% CI Rhat
sd(Intercept)         0.77      0.78        0.02     2.77     1.09
sd(stypeR)            0.94       0.92       0.02     3.22     1.28
cor(Intercept,stypeR) -0.18      0.63       -0.97     0.93     1.09
                      Bulk_ESS    Tail_ESS
sd(Intercept)         124         166
sd(stypeR)            11       19
cor(Intercept,stypeR)  33         247
Population-Level Effects:
                      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept             18.03      0.53    16.86    19.44    1.07   85       28
stypeR                 -2.35      0.74    -3.42    -0.67    1.36    9       12
Family Specific Parameters:
                      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma                 0.89       0.03     0.83     0.96    1.05  199      927
Samples were drawn using sampling(NUTS). For each parameter,
Bulk_ESS and Tail_ESS are effective sample size measures, and
Rhat is the potential scale reduction factor on split chains
(at convergence, Rhat = 1).
Warnmeldungen:
1: Parts of the model have not converged (some Rhats are > 1.05).
Be careful when analysing the results! We recommend running more
iterations and/or setting stronger priors.
2: There were 712 divergent transitions after warmup. Increasing
adapt_delta above 0.8 may help.
See http://mc-stan.org/misc/warnings.html
```

¿Cuáles son las razones? En primer lugar, se debe a la naturaleza de los datos de edad, que son discretos y, por tanto, han perdido su forma naturalmente continua. Las personas no nacen de forma discreta, sino continua, aunque no necesariamente distribuida por igual. Por tanto, se necesitaría un modelo para captar esto. El gráfico muestra cómo la Posterior no capta la separación artificial de la variable `age`. ¿Cuáles son las posibilidades? Por un lado, podemos cambiar el modelo, por ejemplo tratando la edad como un factor y eligiendo un modelo apropiado para predecir las categorías – o podemos intentar devolver a la variable `age` su forma continua sin cambiar la distribución como tal. Podemos hacerlo añadiendo valores aleatorios de una distribución uniforme con valores entre -1 y +1 a cada valor de edad. A diferencia del método anterior con `jitter()`, el espacio entre los enteros se agota completamente, lo que da como resultado una forma mucho más continua. `jitter()` sólo desplaza un poco los números, pero no agota completamente el espacio disponible.

```
# create a more continous age variable... R-Code
age.l <- length(age)
age.enh <- age + runif(age.l,0,1)*sample(c(-1,1),age.l,replace=TRUE)
dats$age.enh <- age.enh
```

La desviación media, es decir, la incertidumbre adicional debida a este procedimiento al nivel de grupo es (véase también la Fig. 6.13)

```
> # comparisons summary statistics
> mean(age-age.jit)
[1] 0.009906947
> mean(age-age.enh)
[1] -0.003388764
> mean(age.jit-age.enh)
[1] -0.01329571
> t(apply(cbind(age,age.jit,age.enh),2,fivenum.wn))
           Min     1st Qu. Median 3rd Qu. Max
age        14.000  15.00   16.000  16.000 19.000
age.jit    13.804  14.94   15.812  16.153 19.198
age.enh    13.032  14.78   15.681  16.655 19.564
```

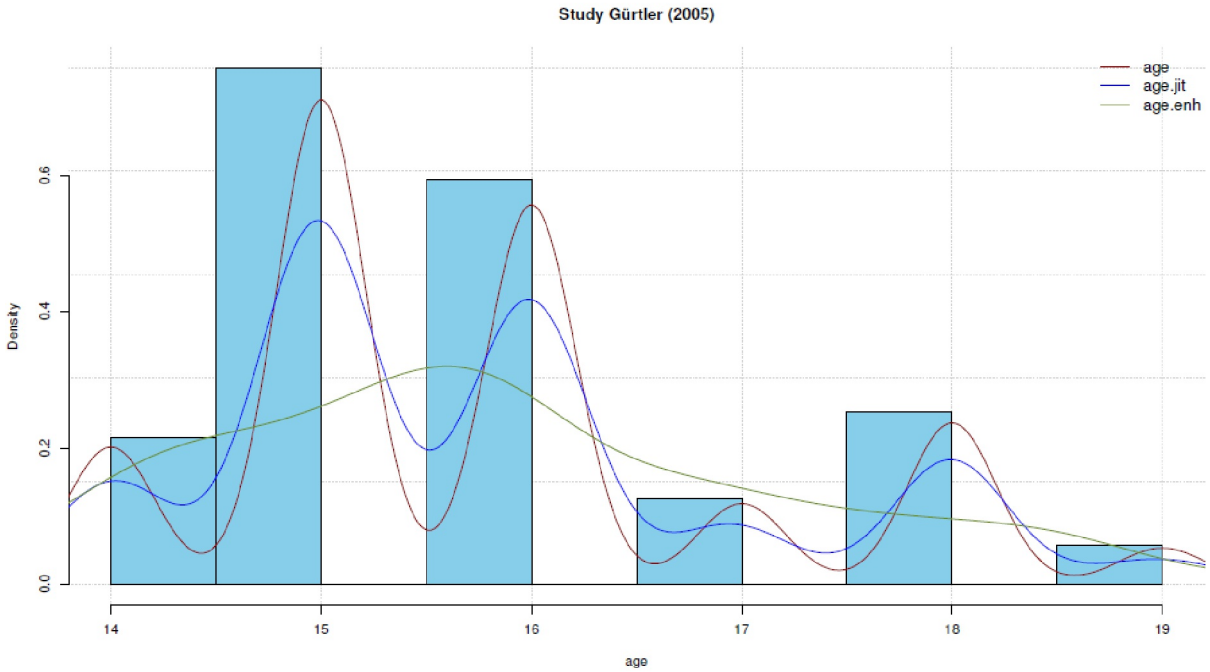


Figura 6.13. *Estudio de Gürtler (2005, edad, categórico, pseudocontinuo)*

No hacemos declaraciones individuales y, por tanto, podemos pasar por alto este factor. Tampoco tenemos en cuenta el hecho de que las personas no nacen distribuidas uniformemente a lo largo del año y que la tasa de natalidad puede variar de un año a otro, es decir, en qué momento nacen más o menos personas. Como sólo interesan las estadísticas de grupo, basta con que éstas permanezcan constantes. Estimemos ahora el mismo modelo con la nueva variable de edad,

```
# that looks pretty well!!!!
brm6 <- brm(age.enh ~ stype, data=dats, family=gaussian(),
save_all_pars=TRUE)
summary(brm6)
plot(brm6)
# much better due to introduction of general uncertainty!
# do not take the uncertainty serious on the level of the single case,
# but in general it is much better marginal_effects(brm6)
plot(marginal_effects(brm6), points=TRUE)
```

las estimaciones de los parámetros resultan ser relativamente similares. Sigue aquí el resumen de `brm6`:

```
> summary(brm6)
Family: gaussian
Links: mu = identity; sigma = identity
Formula: age.enh ~ stype
Data: dats (Number of observations: 317)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
total post-warmup samples = 4000
    Population-Level Effects:
           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept  17.93     0.16      17.62     18.22  1.00 4130     2599
stypeR     -2.48     0.17      -2.80     -2.15  1.00 4155     2909
    Family Specific Parameters:
           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma       1.06     0.04       0.98      1.15  1.00 4125     2991

Samples were drawn using sampling(NUTS).
For each parameter, Bulk_ESS and Tail_ESS are effective sample size
measures, and Rhat is the potential scale reduction factor on split
chains (at convergence, Rhat = 1).
```

Sin embargo, una comprobación del modelo con `pp_check()` ahora no muestra más irregularidades (véase la Fig. 6.14 a la izquierda). El modelo describe los datos de forma muy nítida en comparación con la categorización artificial y poco natural de la edad. Sin embargo, existen diferentes perspectivas. Por ejemplo, el factor Bayes favorece el modelo `brm5`, mientras que, dado el ajuste del modelo, `brm6` parece más sensato (véase la Fig. 6.14 a la derecha, resultado abreviado a continuación):

```
> brm5 <- brm(age ~ stype, data=dats, family=gaussian(),
save_all_pars=TRUE)
> # pp looks better for brm6 compared to brm5
> pp_check(brm5, nsamples=100)
> pp_check(brm6, nsamples=100)
> # but the BF and other criteria show a clear preference
> # for brm5 over brm6
> # ie. the introduced uncertainty is not covered as good
> # as the original data by the factor
> bayes_factor(brm6, brm5)
Iteration: 1
...
Iteration: 4
Estimated Bayes factor in favor of brm6 over brm5: 0.00000
> # Estimated Bayes factor in favor of bridge1 over bridge2: 0.00000
> # not really meaningful, because models are different, ie. outcome
> bayes_factor(brm5, brm6)
Iteration: 1
...
Iteration: 3
Estimated Bayes factor in favor of
brm5 over brm6: 67018420380131365853593600.00000
> # Estimated Bayes factor in favor of
bridge1 over bridge2: 6348672166860855537882842.00000
> # not really meaningful, because models are different, ie. outcome
> # Bayes R^2
> bayes_R2(brm5)
    Estimate Est.Error Q2.5    Q97.5
R2  0.51012  0.027966  0.45046 0.56042
> bayes_R2(brm6)
    Estimate Est.Error Q2.5    Q97.5
R2  0.40317  0.032792  0.33677 0.4615
```

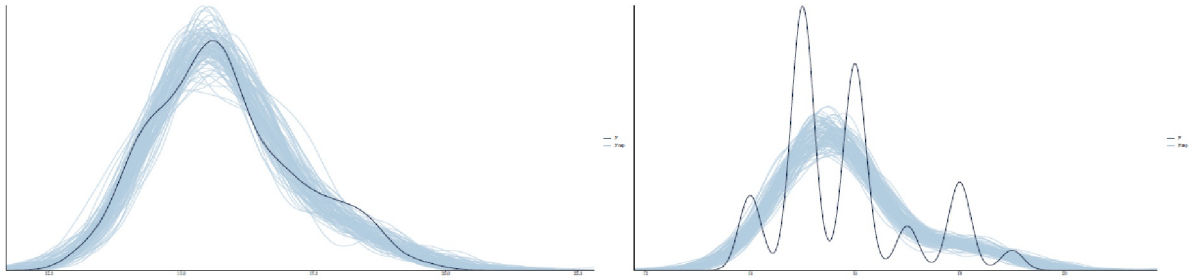


Figura 6.14. *Estudio de Gürtler (2005, comprobación posterior con* `pp_check()`*)*

Si comparamos los coeficientes de los modelos, no parecen tan drásticamente diferentes y cada uno apunta en la misma dirección con estimaciones y errores estándar bastante similares, así como cuantiles:

```
> fixef(brm5)
           Estimate Est.Error Q2.5    Q97.5
Intercept  18.0010   0.13212  17.7408 18.2599
stypeR     -2.5637   0.14285  -2.8453 -2.2817
> fixef(brm6)
           Estimate Est.Error Q2.5    Q97.5
Intercept  17.9266   0.15559  17.6163 18.2210
stypeR     -2.4763   0.16860  -2.7972 -2.1526
> (1-fixef(brm6)/ fixef(brm5))*100
           Estimate Est.Error Q2.5   Q97.5
Intercept   0.41336 -17.759   0.70164 0.21328
stypeR      3.41150 -18.022   1.68873 5.65806
> bayes_factor(brm5, brm6)
Estimated Bayes factor in favor of
brm5 over brm6: 671180327163246048968704 0.00000
```

¿Cuál es ahora el mejor modelo? Puesto que hemos elegido el tipo gaussiano para `brm()`, el ajuste del modelo gráfico con la pseudo-continuidad para la edad parece naturalmente mejor. No debemos tomarnos demasiado en serio el hecho de que el factor de Bayes favorezca claramente a un modelo sobre el otro, ya que el factor de Bayes no es más que una actualización de las expectativas y no ofrece una solución bayesiana completa. El coeficiente de determinación bayesiano R^2 es mayor para el modelo categórico y con un error algo menor, pero tiene un rango desplazado, como puede verse en los cuantiles. Las estimaciones de los propios coeficientes son muy similares. Habría sido mejor disponer de una muestra aún más equilibrada y representativa (para más detalles, véase Gürtler, 2005) y calcular con las fechas de nacimiento para poder utilizar valores continuos reales. Consideramos como causa principal del problema la composición de la muestra a partir de diferentes contextos con grupos de edad no uniformes vemos – y la creación de pseudo-continuidad apoya este punto de vista. La muestra de la escuela secundaria no es representativa. Para ambos tipos de escuela faltan las clases inferiores.

```
> # age vs. stype
> table(age,stype)
stype
age G  R
14  0  34
15  0 120
16  0  94
17  7  13
18 32   8
19  7   2
```

En este sentido, no vemos un modelo como superior al otro, como *supuestamente* sugiere el factor Bayes, sino que vemos ambos desde la perspectiva de cómo surgieron. No debemos olvidar que el modelo es el mismo: `age ~ stype`, `age.jit ~ stype` respectivamente `age.enh ~ stype`; sólo que la variable dependiente es una vez categórica y dos veces pseudocontinua. Esto debería bastar para comprender que contrastar ciegamente los modelos entre sí con un factor de Bayes también puede conducir a graves errores si no se comprende el trasfondo de los modelos. No sólo hemos utilizado un modelo gaussiano (véase el script de R), es decir, con el supuesto de una variable dependiente continua, sino también un modelo categórico de Poisson `brm4`, cuyos resultados no se publican aquí. Sin embargo, las estimaciones se encuentran dentro de un rango comparable entre `brm5` y `brm6`.

Si ampliamos nuestra visión, podemos ver aún más a partir de este pequeño ejemplo. Hemos intentado deliberadamente predecir la edad sólo a partir del conocimiento del tipo de colegio. Si nos fijamos en una muestra representativa imaginada, la Realschule va del 5º al 10º curso, el Gymnasium del 5º al 12º curso (o al 13º en el caso del G13). Así pues, las edades de los cursos 5º a 10º se solapan prácticamente por completo. Además, existe el grupo de edad parcialmente separado y posiblemente muy heterogéneo de los Gymnasien de tarde y los Gymnasien vocacionales. Los grupos de edad pueden ampliarse un poco hacia arriba, dependiendo de la matriculación y del caso de una repetición de curso en la escolarización ordinaria. En este contexto, mencionamos sólo por formalidad que no nos hemos preocupado especialmente por una Prior bien informada que pudiera formular las edades de acuerdo con la distribución de edades anterior entre los tipos de escuela. Con más esfuerzo, un modelo de efectos mixtos que permitiera diferentes pendientes de regresión en función del subgrupo probablemente arrojaría aún más resultados. Pero básicamente, la edad no es suficiente aquí y el género como otra variable popular en contextos sociales apenas añadiría mucho, siempre y cuando la distribución por sexos no difiera especialmente entre Realschule y Gymnasium. Así pues, el coeficiente de determinación R^2 bayesiano parece muy realista y no está nada mal. Con una muestra completa podría incluso disminuir, ya que los solapamientos (véase más arriba) serían entonces mayores. (véase más arriba).

Si realmente queremos describir mejor los datos, necesitamos variables que muestren realmente diferencias de contenido entre los grupos, por ejemplo, datos de rendimiento, posiblemente intereses, etc. Esto requiere entonces una buena hipótesis.

6.8.1.1 Coeficiente bayesiano de determinación R^2

El coeficiente de determinación permite hacer afirmaciones sobre la varianza explicada por el modelo. Desde el punto de vista bayesiano se rechaza el cálculo clásico habitual a partir de los datos predichos divididos por la varianza de los datos, ya que en el caso de las estimaciones bayesianas el numerador de esta fórmula clásica puede llegar a ser mayor que el denominador. Un coeficiente de determinación bayesiano R^2 es el proporcionado por Gelman, Goodrich, Gabry y Vehtari (2019, p.3, fórmula 2, en el código R del apéndice). Se modela siguiendo el procedimiento de los análisis de supervivencia, que utilizan una medida comparable. Técnicamente, se calcula como el cociente de la varianza de los datos pronosticados dividido por la varianza de la suma de los datos pronosticados y la varianza del error, cada una relacionada condicionalmente con los parámetros del modelo.

$$R^2_{Bayes} = \frac{\text{Varianza explicada}}{\text{Varianza explicada} + \text{Varianza residual}} \tag{6.47}$$

Los componentes se calculan ahora con

$$\text{varianza explicada} = \text{varianza de los valores medios predictivos modelizados} \tag{6.48}$$

$$= V_{n=1}^{N} \mathrm{E}(\tilde{y}_n \mid \theta) \tag{6.49}$$
$$= V_{n=1}^{N} y_n^{pred}$$

varianza residual = varianza residual modelizada (6.50)

$$= \mathrm{E}(V_{n=1}^{N}(\tilde{y}_n - y_n^{pred}) \mid \theta) \tag{6.51}$$

$$y_n^{pred} = \mathrm{E}(\tilde{y}_n \mid X_n, \theta) \tag{6.52}$$

V denota la varianza,

E el valor esperado,

N el tamaño de la muestra,

$\tilde{y}_n$ una observación futura (es decir, la predicción de datos futuros aún desconocidos por el modelo) basada en los predictores X_n;

y_n^{pred} denota el predictor lineal y los parámetros estimados del modelo en modelos lineales.

La distribución posterior de θ genera la distribución predictiva posterior de y_n^{pred} (véase también el capítulo 6.8.4.3).

Por lo tanto, R^2 también es condicional y toma valores entre cero y uno, independientemente del modelado de ypredn . En R, este coeficiente puede calcularse con `bayes_R2()` del paquete R `brms.` Sin embargo, también es posible con

```
brm.fit <- add_criterion(brm.fit, criterion=c("loo","waic","kfold","R2","marglik"))
```

ya que pueden añadirse diferentes criterios de calidad a los modelos ajustados con `brm()`. Una comparación de este coeficiente de determinación R^2 en la variante bayesiana muestra lo siguiente para los modelos en función de la modelización de la edad (`ptII_quan_Bayes_BayesFactors_test-hypos.r`): en primer lugar que no podemos compararlos directamente, ya que por un lado se trata de una variable dependiente categórica y por otro de una variable dependiente pseudocontinua y, en sentido estricto, se trata de variables diferentes.

```
# compare brm5 and brm6 -> differences for sigma, not fixed effects...
brm5 <- add_criterion(brm5,
        criterion=c("loo","waic","kfold","R2","marglik"))
brm6 <- add_criterion(brm6,
        criterion=c("loo","waic","kfold","R2","marglik"))
```

Si lo comparamos, hay advertencias:

```
> # should work if not brm5 and brm6
# would have a different dependent variable
> for(i in c("loo","waic","kfold"))
+ {
+ cat("\n",i)
+ print(loo_compare.brmsfit(brm5,brm6, criterion=i))
+ }
```

```
loo elpd_diff se_diff
brm5 0.0 0.0
brm6 -57.3 10.3
waic elpd_diff se_diff
brm5 0.0 0.0
brm6 -57.3 10.3
kfold elpd_diff se_diff
brm5 0.0 0.0
brm6 -57.2 10.4
Warnmeldungen:
1: Not all models have the same y variable.
('yhash' attributes do not match)
2: Not all models have the same y variable.
('yhash' attributes do not match)
3: Not all models have the same y variable.
('yhash' attributes do not match)
```

Ahora podemos seguir tomando un modelo categórico en lugar de uno continuo para la edad y cambiar a la distribución de Poisson para predecir las frecuencias de palabras. Para ello, los datos necesitan algún formato, que se puede hacer con `ftable()`.

```
# add counts instead of age -> R-Code
# agecounts + vars (counts) + log(...) + (1|stype)
# prepare data with ftable()
ftable(age ~ sex,data=dats)
ftable(age ~ sex + stype,data=dats)
count.age <- as.data.frame(ftable(age ~ sex + stype,data=dats))
colnames(count.age)[colnames(count.age) == "Freq"] <- "Freq.age.cat"
count.age
```

Ya no podemos comparar los coeficientes directamente porque ahora la variable dependiente está definida de forma diferente y los predictores también están parcialmente definidos.

```
summary(brm.p1.0)
```

En primer lugar, podemos examinar las Posteriores y el ajuste del modelo y compararlos con `brm6` (véase la Fig. 6.15):

```
plot(brm.p1.0)
pp_check(brm.p1.0)
# plot effects
plot(marginal_effects(brm.p1.0), points=TRUE)
```

y luego comparar las explicaciones de varianza con `bayes_R2()` y los modelos entre sí con `bayes_factor()`.

```
> bayes_R2(brm.p1.0)
Estimate Est.Error Q2.5 Q97.5
R2 0.74679 0.047434 0.63991 0.82425
> # BF
> bayes_factor(brm.p1.0, brm6)
Iteration: 1
Iteration: 2
Iteration: 3
Iteration: 4
Iteration: 5
Estimated Bayes factor in favor of brm.p1.0 over brm6: 2.393324117e+143
```

El ajuste posterior y el ajuste del modelo parecen limpios. Por comparación, el modelo de Poisson `brm.p1.0` se prefiere sistemáticamente a `brm6`. Las pruebas de comparación con otros criterios de calidad como `loo()` fallan, ya que estamos examinando el mismo contenido, pero estadísticamente muy diferente (tamaño de la muestra, definición de predictores y variable dependiente, etc.).

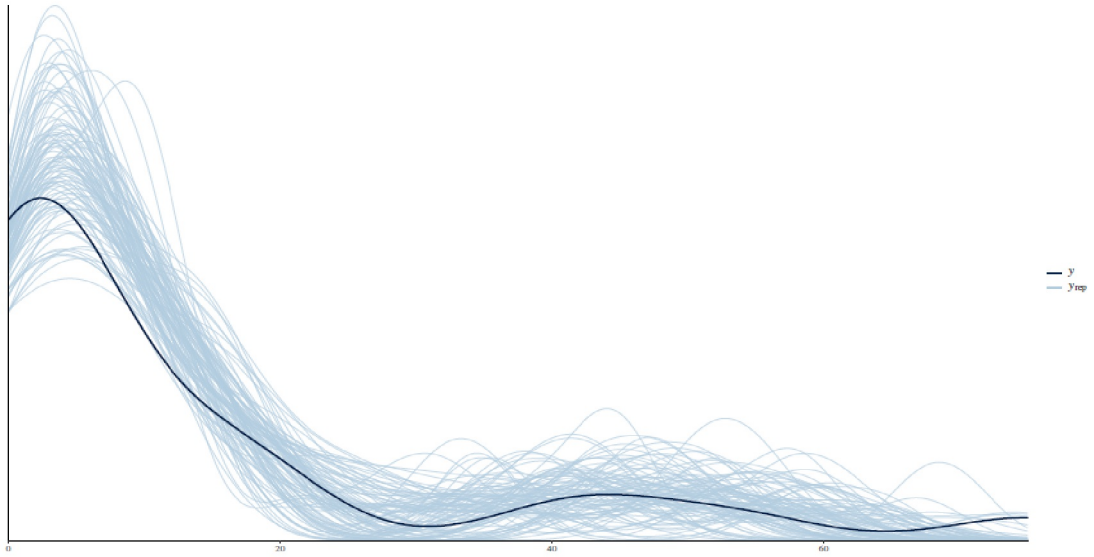


Figura 6.15. *Estudio de Gürtler*

(2005, edad clasificada categóricamente, comprobación posterior con `pp_check()`).

```
> # does not work, because models are too different
> loo(brm6, brm.p1.0, reloo=TRUE)
Fehler: Models have different number of observations.
```

Podemos hacerlo si los modelos difieren sólo en sus predictores, por ejemplo.

¿Qué aprendemos de esto? Los factores de Bayes son útiles para evaluar críticamente la inclusión de nuevas variables en un modelo. Sin embargo, son de poca ayuda cuando se trata de la adecuación general del modelo. En este caso, son las Posteriores y los métodos analíticos asociados, como las comprobaciones predictivas posteriores, el trazado de valores posteriores, etc. En lo que respecta a los factores de Bayes, según las explicaciones anteriores la Prior elegida tiene una posición significativa. Según Rouder, Speckman, Sun, Morey e Iverson (2009), la distribución Cauchy se implementa de forma estándar en el paquete `BayesFactor` de R. Según una entrada de blog de Richard Morey, el autor de `BayesFactor`, una Prior siempre debe estar justificada (Morey, 2014), ya que elabora un ejemplo de datos y muestra que la elección de la Prior no es necesariamente única, sino que debe elegirse en función de las necesidades,

> „The Cauchy prior was suggested by Rouder et al. (2009), and is the one implemented in the BayesFactor package. Of course, this half-Cauchy distribution is not the only way we could implement Paul's hypothesis.We could choose a di erent shape of distribution; we could make it wider or narrower around 0; we could even restrict the range further.We must choose some valid probability distribution in order to create a hypothesis that constrains the data, and we should do our best to make the test meaningful by choosing a defensible distribution."

En consecuencia, existe una gran variedad de publicaciones sobre el tema de un grupo de autores que intentan justificar la elección de la Prior en el contexto de los factores de Bayes o de la Prior misma (entre otros, Rouder, Speckman, Sun, Morey e Iverson, 2009; Morey y Rouder, 2011; Rouder, Morey, Speckman

y Province, 2012; Rouder y Morey, 2012; Ly, Verhagen y Wagenmakers, 2016; Romeijn,Morey y Rouder, 2016).

Tarea 6.4: Perfeccionamiento del modelo

El modelo anterior puede refinarse aún más a continuación, lo que constituye una tarea para los lectores. Por ejemplo, se podría introducir el género como predictor, examinar los datos en busca de posibles interacciones o considerar si tiene sentido ejecutar un HLM en lugar de un modelo lineal simple. Las posibles variantes serían

```
Freq.age.cat ~ age + stype R-Notation
Freq.age.cat ~ age + stype + sex
Freq.age.cat ~ age + stype * sex
Freq.age.cat ~ (1|age) + stype * sex
```

Esto se puede seguir con gráficos de interacción para encontrar hipótesis adecuadas de forma exploratoria:

```
# interaction plots R-Code
colos <- colorRampPalette(c("blue","darkred"))
with(count.age, interaction.plot(sex, stype,
       Freq.age.cat, legend=TRUE, col=colos(2)))
```

El factor de Bayes expresa una *suposición sobre la plausibilidad relativa de las hipótesis en comparación directa*, cada una de las cuales es un candidato potencial para el proceso de generación de datos. No se trata de la Posterior, sino del cambio en la expectativa (Prior) debido a los datos (Likelihood). La pregunta es, qué hipótesis explica mejor los datos y para ello se compara el cociente de las probabilidades de los datos para la validez de cada una de las hipótesis en liza. Esto puede ser útil a la hora de tomar decisiones dados dos o más modelos en competencia. Por otra parte, opera entonces un principio de exclusión y no de integración. No se trata de ambos y en qué condiciones, sino de un *o bien* sumario. Queda por ver si esto siempre tiene sentido, ya que este sentido *o bien* está anclado estructuralmente en el análisis y no puede leerse a partir de los coeficientes numéricos resultantes. No se trata de una cuestión matemática, sino de un problema de diseño y epistemológico. El problema ya existe antes de que cualquier tipo de análisis de datos. Kass y Raftery (1995, p.788) expresan muy bien esta discrepancia en su intento de justificar por qué comprueban estrictamente las hipótesis.

> „We introduced Bayes factors as a way of assessing the evidence in favor of a scientific theory. Our statement that Jeffrey's approach computes 'the posterior probability that one of the theories is correct' invites argument, however. Some would say that theories are never correct, and thus any approach that assumes they are must be flawed. Those who make this argument generally prefer the use of interval estimates."

Esto lo resume bien: en última instancia, es responsabilidad del investigador individual cómo construye su diseño y qué pregunta de investigación está tratando. Las conclusiones sólo pueden hacerse en el marco del diseño. Los factores de Bayes son la aplicación de un principio de selección no cooperativo y estrictamente evolutivo. En la evolución, sin embargo, dependiendo de las condiciones, sobrevive el que aprende a cooperar. Dado que la cooperación no puede estudiarse experimentalmente debido a los largos periodos de tiempo que implica, los experimentos suelen llevarse a cabo en el contexto de la teoría de juegos (por ejemplo, Peters & Adamou, 2015). Con los factores de Bayes las hipótesis se contrastan entre sí y se rechazan o mantienen según de Popper. No cooperan entre sí, sino funcionan según el principio económico de quién puede superar a mercado frente a los demás competidores. Algo parecido se practica en el contexto del análisis de secuencias en el que tampoco se trata de permitir que existan en paralelo todas las posibles

hipótesis de estructura de casos, sino de generar una hipótesis de estructura de casos integrada que pueda explicar todo lo posible, por un lado, y sea superior a las demás hipótesis competidoras, por otro. No se obtienen ganancias absolutas, sino sólo relativas. Esto es similar a los hallazgos de las ciencias sociales (Stanger, Jhangiani & Tarry, 2014, cap. 12) de que las personas tienden a ganar en términos relativos a los demás, pero pueden perder en términos absolutos -- lo cual no es muy conveniente.

Aquí, sin embargo, el proceso de hipótesis de estructura de casos resulta ser una mezcla de ajuste y prueba de modelos, ya que, por un lado, se descartan o retienen hipótesis y, por otro, el modelo que emerge lentamente se ajusta virtualmente a través de la condensación y la nueva información. Si se considera el proceso de ajuste del modelo y no la mera comprobación, estos procesos son muy muy similares, de modo que el análisis de secuencias parece más bien un complejo ajuste cualitativo de modelos, pero basado en un largo proceso de generación de hipótesis, comprobación y rechazo según los criterios de falsación de Popper.

Un examen más detallado de la ecuación del factor de Bayes (véase la Ec. 6.42, p.541) también muestra que el BF_{21} es el cociente de dos Likelihoods. Si ambas hipótesis o modelos se consideran igualmente probables, de modo que $Prior(H_2) = Prior(H_1)$, la ecuación se reduce al cociente de las dos distribuciones posteriores.

$$\mathrm{BF}_{21} = \frac{p(H_2 \mid D)}{p(H_1 \mid D)} \tag{6.53}$$

Tabla 6.8: *Función de pérdida y mejores estimadores*

tipo	función de pérdida	mejor estimador
binario	$(0, 1)$	Moda
lineal	$\lvert d - p \rvert$	Mediana
cuadrático	$(d - p)^2$	Media

Otra conexión con la estadística clásica es que si las probabilidades de la ecuación BF_{21} se sustituyen por las respectivas estimaciones de máxima Likelihood, la prueba se transforma en una prueba clásica de Likelihood Ratio (LRT). Al integrar todo el espacio de parámetros (para cada modelo), el factor de Bayes es independiente de los parámetros individuales, lo que no ocurre con la LRT. La ventaja del BF_{21} es que se penalizan demasiados parámetros del modelo, lo que protege contra el sobreajuste (véase más adelante, Kass & Rafter, 1995). Una circunstancia también conocida en la estadística clásica (por ejemplo, la estimación mediante máxima Likelihood completa frente a máxima Likelihood restringida (REML, Pinheiro & Bates, 2009). Sin embargo, un enfoque sólo tiene sentido si también se sabe cómo es el beneficio en relación con el esfuerzo, modelable mediante funciones de pérdida.

6.8.1.2 Funciones de pérdida

Los factores de Bayes suelen estar provistos de una función de pérdida (Davidson-Pilon, Capítulo 5) para estimar los costes de una decisión (inferencia) errónea (Gelman, Carlin, Stern & Rubin, 2004, Capítulo 9). Hennig y Kutlukaya (2007) analizan el diseño básico de una función de pérdida. A partir de la combinación de una función de pérdida con un enfoque de Bayes con un factor de Bayes, en principio puede derivarse una decisión razonada basada en el contexto. Esto parece más que simpático, ya que las ciencias sociales en particular rara vez prestan atención a si el esfuerzo necesario para generar conocimiento o las consecuencias del proceso de generación de conocimiento son beneficiosas para la humanidad, según criterios estrictos que pueden o incluso deben incluir factores económicos. Esta perspectiva adicional puede crear nuevos conocimientos, siempre y cuando no se utilice indebidamente para controlar y dirigir la ciencia por

organismos externos. No hay que olvidar que debe y no debe haber ámbitos que puedan ponerse simplemente en función de las pérdidas, por ejemplo, cuando se trata de invertir en algo nuevo a largo plazo y esto no conduce inmediatamente a resultados aprovechables. Puede que no sea muy sensato detener todas las inversiones sólo porque algo es complejo y difícil o porque los avances llevan tiempo. Muchos de los grandes inventos de la humanidad no se desarrollaron mediante planes maestros con control de costes, pero fueron con un trasfondo económico. En este sentido, el uso de funciones de pérdidas tiene sentido, pero, como siempre, requiere un razonamiento que no se centre únicamente en razones externas mundanas como el ahorro de personal y dinero.

McElreath (2015, p.59s.), Davidson-Pilon (2015), Clyde, Cetinkaya-Rundel, Rundel, Banks, Chai y Huang (2019-03-29, cap. 3), Bååth (2014, 2015) y Neto (2015) ofrecen ejemplos sencillos de una función de pérdidas, algunos de ellos con código R. Hay varias formas (véase la Tabla 6.8 para una selección) de detectar desviaciones de un valor óptimo y minimizar la pérdida, que a la inversa equivale a la maximización del beneficio. Una pérdida binaria (0; 1) asigna la moda como el mejor estimador, mientras que la diferencia absoluta $|d - p|$ conduce a la mediana posterior, que divide la distribución posterior al 50% de la masa en dos partes iguales. La mediana posterior es, por tanto, una estimación puntual del valor que minimiza la desviación absoluta. Diferentes funciones de pérdida conducen a diferentes estimaciones puntuales de estos valores óptimos. Una función de pérdida cuadrática $(d - p)^2$ conduce a la media de la distribución posterior. Si la distribución posterior tiene una forma simétrica similar a la distribución normal, la moda, la mediana y la media son aproximadamente iguales. Dependiendo de la desviación de esta forma normal y de la asimetría de la distribución posterior, estos puntos se encuentran separados entre sí. Una función de pérdida que trabaja con una potencia (por ejemplo, dicha función de pérdida cuadrática) es naturalmente sensible a los valores atípicos y a los valores muy grandes. Sin embargo, esto no significa que no sea más apropiada en un contexto particular que si el exponente es sólo 1. La elección de la función de pérdida no es más definitiva y absoluta que la elección de la Prior. Tampoco es necesario elegir la distancia directa como en los ejemplos $|d - p|$ o $(d - p)^2$, sino que la función puede definirse de forma completamente diferente para distintos rangos de valores e incluir muchos otros valores y parámetros en el cálculo. Los ejemplos anteriores sólo representan la "variante" más sencilla. La diferencia simplemente define la relación. Esto significa que una pérdida cambia proporcionalmente a la distancia desde el valor óptimo – linealmente, cuadráticamente, etc. Toda la relación puede ser igualmente no lineal e incluir umbrales naturales a partir de los cuales, por ejemplo, debe iniciarse una determinada acción (por ejemplo, aumentar el personal porque la masa de pedidos ya no se pueden gestionar).

6.8.1.3 Caso práctico de función de pérdidas

He aquí un ejemplo (`ptII_quan_Bayes_lossfun_startagain.r`) del paquete R y R `LaplacesDemon`:

```
?lossMatrix
```

Creemos un *escenario ficticio* a partir de esto en el contexto del estudio repetidamente mencionado de Studer (1995, 1998) sobre las tasas de éxito en la terapia de adicción para pacientes hospitalizados. Así, podríamos conjeturar que el cantón de Zúrich (Suiza), en el marco de determinadas tasas de éxito – operacionalizadas como tasas de finalización con éxito, es decir, que los clientes terminan el programa completo y no lo abandonan – hace diferentes contribuciones a la financiación. Sin embargo, hay un límite superior por encima del cual no se paga nada proporcionalmente y también un límite inferior por debajo del cual no se paga nada en absoluto, sino que se cuestiona la calidad del trabajo – con la amenaza de retirada de la certificación o la licencia. Estos valores pueden variar de un año a otro y se basan en las tasas generales de éxito de los centros de tratamiento de adicciones en régimen de internado en los países de habla alemana. De ahí se pueden derivar los siguientes límites: elegimos valores ficticios:

- Por debajo del 20% de aprobados, se suspenden los pagos y peligra la certificación del centro. Esto pone en peligro la supervivencia económica de la institución.

- Entre el 20% de tasa de aprobados, hay una financiación básica de 450 CHF/día. El Cantón de Zúrich fija el porcentaje de aprobados típico en un 43% en un año. Este porcentaje puede variar de un año a otro y depende también de las demás instituciones. Todo lo que supere este índice recibe una bonificación proporcional a la distancia de este óptimo: cuanto mayor sea el índice de aprobación, mayor será la bonificación. Sin embargo, la bonificación es de un máximo del 20% por tasa diaria comprendida entre el 43%. Así pues, la tasa máxima es de 450 *1.2 = 540 CHF/día.
- Por encima del 60% de tasa de aprobados, no se concede ninguna otra bonificación, al menos no monetaria. Sin embargo, la institución se considera de referencia, lo que puede tener otras ventajas en términos de publicidad, marketing, captación de fondos de terceros y donaciones, etc.

A partir de ahí, se podría obtener una estimación de la función de pérdidas o beneficios anuales. Si las cifras cambian, lo cual es realista, la función de pérdidas cambia. El resultado es el desglose de la tabla 6.9.

Tabla 6.9: *Función de beneficio/pérdida*

p_{low}	p_{up}	Beneficio(CHF)	Decisión/ Consecuencia	monetaria
0	0.2	0	Retirada de la certificación/ Licencia	0
0.2	0.43	450	Financiación básica	450
0.43	0.6	450 + Bonificación	Bonif. B = $p_{diff} \cdot 450$ (Máx = $450 \cdot 1.2$)	450 a 540
0.6	1	450 + Bonificación	acceptado como institución referencial	540

La figura 6.17 contiene ahora varios tipos de funciónes de pérdida concebibles. Como puede verse en desarrollo de una Posterior utilizando el ejemplo de las tasas de aprobados en terapia de adicciones (véase el capítulo 6.16 o 6.15.2), la función de pérdida tendría un aspecto diferente según el año examinado. Trabajamos con la función de pérdida a lo largo de todos los años (1992-2017, véase la tabla 6.14). Observamos la pérdida simple y al cuadrado tomando las desviaciones simple y al cuadrado, respectivamente, para cada valor de 0 a 1 y ponderando – es decir, multiplicando – por la Posterior. Esto da como resultado los dos gráficos superiores de la Figura 6.17, para la pérdida lineal simple y para la pérdida cuadrática. En los gráficos, cada punto marca una función de pérdida distinta. El mínimo es fácil de leer. La minimización de la pérdida simple conduce a la mediana de la Posterior, la de la pérdida cuadrática a su media (McElreath, 2015, p.60) y una pérdida binaria se refiere a la moda de la Posterior (véase la Tabla 6.9). Los gráficos reflejan la desviación respecto a un valor de referencia. Para ello, no elegimos el máximo de la Posterior, sino el valor de referencia externo ficticio del cantón de Zúrich, es decir, p = 0.43. Si este valor de referencia representara el óptimo económico de la institución, una desviación al alza o a la baja significaría cada una pérdida relativa. Una desviación a la baja significa entonces quedarse por debajo de las propias posibilidades y posiblemente tener muy pocos clientes para poder trabajar económicamente. Una desviación al alza significa un exceso de inversión (por ejemplo, personal, material, otros costes), es decir, que las inversiones superan los ingresos reales. Aunque la tasa de rendimiento está entonces por encima de la referencia, el beneficio económico no está claro. Esto no contrasta con el hecho de que la tasa de rendimiento más alta posible es esencial como punto de partida para una recuperación satisfactoria. Pero desde una perspectiva económica, es necesario considerar qué costes conlleva. Es posible que a partir de un determinado índice de aprobados, los costes adicionales aumenten enormemente, mientras que el aumento simultáneo del índice de aprobados se queda atrás en comparación. Esto lleva a un debate ético que, como es lógico, no se corresponde necesariamente con la perspectiva económica. Así pues, los dos gráficos muestran el efecto de una desviación de una referencia. Por supuesto, podríamos razonar y elegir una referencia completamente distinta.

Para el cálculo, partimos de los porcentajes de aprobados y estimamos o trazamos la correspondiente posterior (ver Fig. 6.16) con R (`ptII_quan_Bayes_lossfun_startagain.r`).

```
# set directory
sa.3 <- read.table("startagain_statistics_1992-2017_all-out.tab",
+ header=TRUE, sep="\t")
head(sa.3)
steps <- 1000
theta <- seq(0,1,length.out=steps)
sa.3.d <- dim(sa.3)
pbl.res <- exp(pbl(theta=theta, si=sa.3[sa.3.d[1],"s.cs"],
+ Ni=sa.3[sa.3.d[1],"N.cs"], loga=TRUE))
pjc.res <- exp(pjc(theta=theta, si=sa.3[sa.3.d[1],"s.cs"],
+ Ni=sa.3[sa.3.d[1],"N.cs"], loga=TRUE))
# plot
sN.ME.res <- data.frame(pbl.res, pjc.res)
plot.bl.jc(theta, sN.ME.res=sN.ME.res, si=si, Ni=Ni,
+ filling=TRUE, sele=c(0.4,0.6))
theta.mat <- data.frame(theta, pbl.res)
# we work with the pbl and not pjc, difference
# between both is not really much if at all
theta.mat.pjc <- data.frame(theta, pjc.res)
colnames(theta.mat) <- c("theta","post")
# MAP = maximum aposteriori
theta.map <- theta.mat[theta.mat[,"post"] ==
+ max(theta.mat[,"post"]),"theta"]
# mean
theta.mean <- sum(theta.mat[,"theta"] * theta.mat[,"post"] / steps)
```

O como alternativa

```
R-Code # alternative from McElreath p.60f.
# median = min linear loss function
# infids <- apply(res.Xct,2, function(i)
which(is.infinite(i),arr.ind=TRUE))
# theta.mat <- res.Xct[-infids$pbl.res,]
theta.linear <- sapply(theta.mat[,"theta"], function(x)
{
sum(theta.mat[,"post"]*abs(x - theta.mat[,"theta"]))
}
theta.median <- theta.mat[which.min(sapply(theta.mat[,"theta"],
function(x) sum(theta.mat[,"post"]*abs(x - theta.mat[,"theta"]))
)) ,1]
# mean = min quadratic loss function
# MW <- res[which.min(sapply(res[,1],
function(x) sum(res[,2]*abs(x - res[,1])^2)
)) ,1]
theta.quad <- sapply(theta.mat[,"theta"],
function(x) sum(theta.mat[,"post"]*abs(x -
theta.mat[,"theta"])^2))
theta.mean <- theta.mat[which.min(sapply(theta.mat[,"theta"],
function(x) sum(theta.mat[,"post"]*abs(x -
theta.mat[,"theta"])^2 ))) ,1]
```

A partir de ahí se pueden obtener las estadísticas sumarias.

```
> # theta mean = a/(a+b)
> theta.map
[1] 0.4955
> theta.median
[1] 0.4955
> theta.mean
[1] 0.4955
```

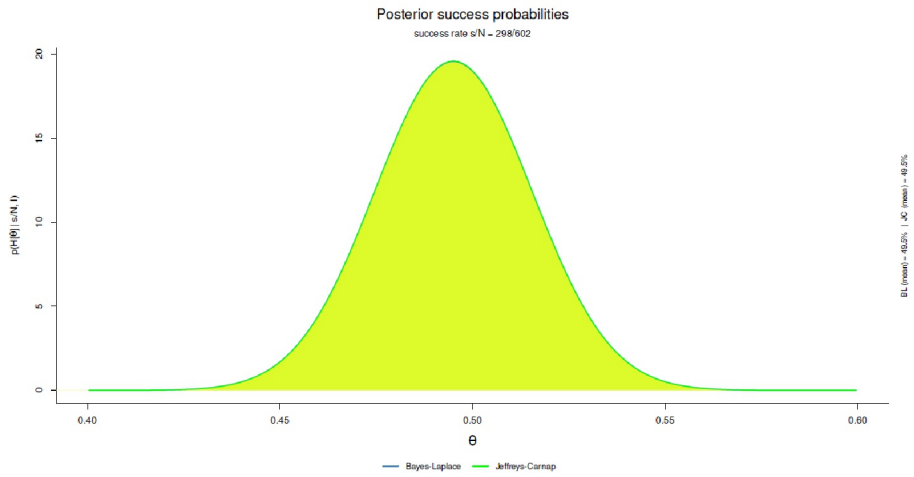


Figura 6.16. *start again 1997-2017 (índices de aprobados, Posterior)*

Ahora se fija el valor de referencia y una función calcula para cada Posterior la pérdida como desviación lineal o cuadrática del valor de referencia. A continuación, se representa el resultado gráficamente. Empezamos con la desviación lineal

```
# actual loss function calculation
# define reference
ref <- 0.43
# calculate deviations for the loss function
# linear
theta.linear <- sapply(theta.mat[,"theta"],
function(x) sum(theta.mat[,"post"]*abs(x - ref)))
```

y añadimos la desviación cuadrática:

```
# quadratic
theta.quad <- sapply(theta.mat[,"theta"],
function(x) sum(theta.mat[,"post"]*(x - ref)^2 ))
```

Los dos gráficos inferiores siguen las directrices de la Tabla 6.9. El gráfico inferior izquierdo muestra cómo resulta un beneficio dentro de un rango limitado. Por encima y por debajo de los umbrales críticos del 20% y el 60%, se produce una pérdida total o ningún beneficio adicional (lo que aquí se define como pérdida). En medio se encuentra la financiación básica. El gráfico inferior derecho añade el sistema de bonificación, que se detiene por encima del 60%. Desde un punto de vista económico, está claro en qué rango debe operar la instalación: entre el 43% y el 60%, y si el índice de aprobados actual es del 49.74%, esto sería motivo para continuar con la institución exactamente como está o para seguir aumentando la calidad.

Las funciones de pérdida, que parten de un rango limitado y añaden el sistema de bonificaciones, también se calculan cada una mediante θ y se representan gráficamente. En primer lugar, la función para la zona limitada en la que invertir

```
# limited area where to invest
# specific context
theta.linear1 <- sapply(theta.mat[,"theta"],
function(x) mean(theta.mat[,"post"] *
(abs(x - ref) > 0 & abs(x-ref) < .17)))
```

así como el sistema de bonificación, una vez como zona limitada y otra con el límite monetario hacia arriba

```
# limited area to invest + bonus
theta.linear2 <- sapply(theta.mat[,"theta"], function(x) {
DIFF <- x-ref
BENEFIT <- 0
if(abs(DIFF) > 0 & abs(DIFF) < maxdist) BENEFIT <- 1
if(DIFF > 0 & abs(DIFF) < .17) BENEFIT <- 1 + DIFF
mean(theta.mat[,"post"] * BENEFIT)
} )
# bonus and upper limit
uplimit <- 0.6
theta.b <- seq(ref,uplimit,0.01)
satz <- 450
fac <- 1.05
bonus <- satz * seq(1,1.05,length.out=18)
theta.l <- length(theta.b[theta.b >= 0.43 & theta.b < 0.6])
theta.b <- theta
theta.b[theta.b < 0.2] <- 0
theta.b[theta.b >= 0.2 & theta.b < 0.43] <- 450
theta.b[theta.b >= 0.43 & theta.b < 0.6] <- satz *
seq(1,1.05,length.out=theta.l)
theta.b[theta.b >= 0.6 & theta.b <= 1] <- satz*fac
```

y ahora los gráficos (váse fig. 6.17), que visualizan todo:

```
# everything on one plot... R-Code
par(oma=c(2,1,4,1), "cex.axis"=1, bty="l", mfrow=c(3,2))
# linear loss
plot(theta.mat[,"theta"], theta.linear, main="linear loss",
       type="l", bty="n", xlab=expression(theta), col="darkred",
       ylab="linear loss", pre.plot=grid(), cex.lab=1.2)
abline(v=ref, lty=2, lwd=2, col="blue")
text(ref,max(theta.linear)/2, "ref", col="blue", pos=2)
# quadratic loss
plot(theta.mat[,"theta"], theta.quad, main="quadratic loss",
       type="l", bty="n", xlab=expression(theta), col="darkred",
       ylab="quadratic loss", pre.plot=grid(), cex.lab=1.2)
abline(v=ref, lty=2, lwd=2, col="blue")
text(ref,max(theta.quad)/2, "ref", col="blue", pos=2)
# linear1 limited area to invest
ylim <- c(0,1.2)
plot(theta.mat[,"theta"], theta.linear1, ylim=ylim,
main="linear loss functions (with limited area to invest)",
       type="l", bty="n", xlab=expression(theta), col="darkred",
       ylab="linear loss + investment area", pre.plot=grid())
abline(v=theta.median, lty=2, lwd=1, col="blue")
abline(v=theta.mean, lty=2, lwd=1, col="red")
abline(v=ref, lty=2, lwd=1, col="green")
text(theta.median,max(theta.linear1)/2, "median", col="blue", pos=4)
text(theta.mean,max(theta.quad)/2, "mean", col="red", pos=2)
text(ref,max(theta.linear1)/2, "ref", col="green", pos=2)
# linear2 limited area to invest + bonus
plot(theta.mat[,"theta"], theta.linear2,
       main="linear loss (with limited area to invest + bonus)",
       ylim=ylim, type="l", bty="n", xlab=expression(theta),
       col="darkred", ylab="linear loss + area to invest + bonus",
       pre.plot=grid(), cex.lab=1.2)
abline(v=ref, lty=2, lwd=2, col="blue")
text(ref,max(theta.linear2)/2, "ref", col="blue", pos=2)
```

```
# bonus with upper limit
plot(theta, theta.b, pre.plot=grid(), cex.lab=1.2
       main="linear loss (with bonus and upper limit)",
       ylim=c(0,600), type="l", bty="n", xlab=expression(theta),
       col="darkred", ylab="linear loss + bonus + upper limit")
abline(v=ref, lty=2, lwd=2, col="blue")
text(ref,mean(theta.b)/2, "ref", col="blue", pos=2)
# main title
mtext("Loss functions", outer=TRUE, line=0.5, cex=1.8, side=3)
```

Es importante señalar que, aunque matemáticamente exista un mínimo de la función de pérdida, esto no significa que exista en la realidad. El ejemplo pretende demostrar que existen rangos dentro de los cuales se puede registrar una ganancia o pérdida similar y que existen rangos absolutos (por ejemplo, retirada de la certificación) y rangos dentro de los cuales determinados cambios en los valores empíricos sólo tienen consecuencias limitadas.

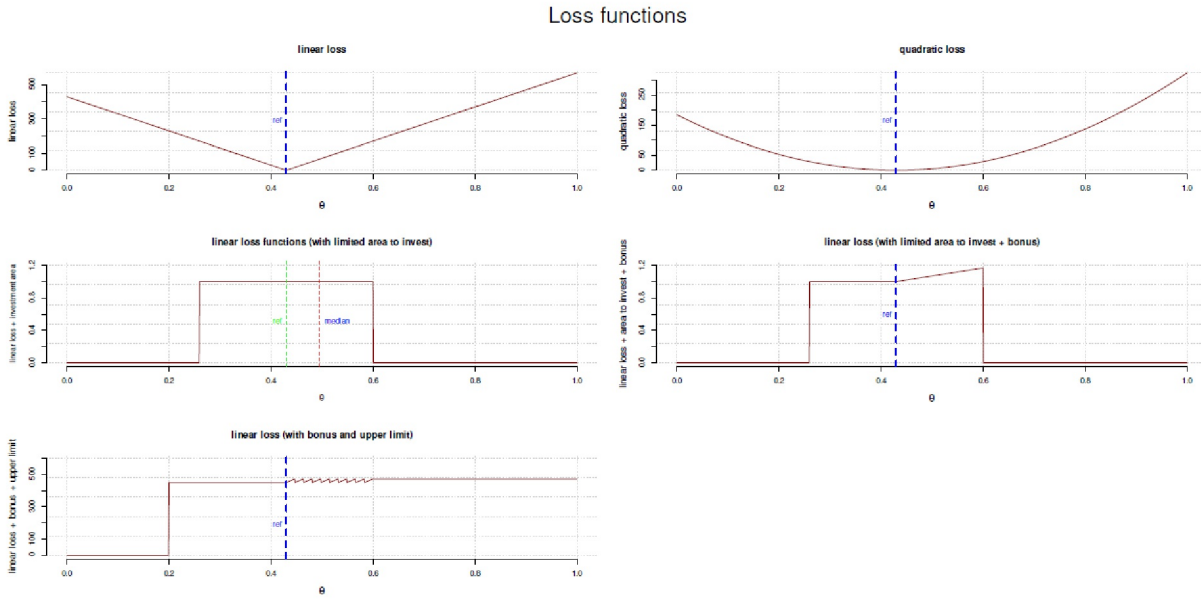


Figura 6.17: *start again (c)*

Son concebibles otras funciones de pérdida. En este sentido, no existe la mejor función de pérdida, sino que establece según criterios relacionados con el contenido y se investiga cómo cambian los datos a lo largo de estos a lo largo de estos criterios, ya que la propia función de pérdida denota lo que es bueno o malo (Hennig & Kutlukaya, 2007).

Puede existir una función de pérdida (comparable a la Tab. 6.9) en forma de matriz, lo que sería apropiado en el caso de las observaciones anteriores sobre la política social del cantón de Zúrich. El paquete R `LaplacesDemon` ofrece posibilidades de análisis para este fin con la función R `LaplacesDemon()` con el fin de tomar decisiones razonables. Para decisiones en contextos que no son predecibles, hay que realizar un trabajo previo para el proceso de toma de decisiones:

- Operacionalización de la pérdida o ganancia esperada y aplicación empírica en forma de datos o análisis de datos.
- Asignación de la probabilidad esperada a priori de las distintas opciones de decisión
- Derivación de las pérdidas o ganancias posteriores a lo largo de una función de pérdida por definir.
- Trazar gráficos y derivar consecuencias

Si bien dicha matriz de pérdidas puede realizarse como una estimación puntual, también puede calcularse como una distribución de valores, de modo que alrededor de las respectivas estimaciones puntuales se añada una zona de incertidumbre más o menos probable, que debería corresponder con las condiciones conocidas en la realidad. Un ejemplo típico de esto sería la inversión en acciones, que en pocas palabras suele significar: las acciones inciertas prometen mayores beneficios, pero también mayores pérdidas. En cambio, las acciones más seguras dan menos beneficios, pero sus pérdidas son más manejables. La cuestión ahora es, ¿en qué acciones hay que invertir? Dado que uno no debe guiarse por la codicia o el puro pensamiento de seguridad, una matriz de decisión de este tipo puede ayudar a tomar una decisión razonable a lo largo de la información contextual disponible, por ejemplo, combinada con información sobre las empresas consideradas. Otra decisión extrema sería evaluar los riesgos de una erupción volcánica inminente.

¿Hay que evacuar o no una ciudad o una zona? El coste de la evacuación es alto, pero también lo es el coste (vidas humanas...) si un volcán entrara en erupción y no se evacuara a la población. Aunque desde el punto de vista geológico existen pruebas de una erupción segura o de la no actividad actual de una erupción, existe una gran zona gris entre medias.

Existe una zona gris, por lo que una decisión con consecuencias tan trascendentales como una evacuación debe elegirse cuidadosamente. En este caso, una matriz de decisión bien fundamentada junto con los datos geológicos disponibles puede ayudar a respaldar una decisión a nivel interno y de cara al público. En consecuencia, se pueden crear muchos ejemplos del mundo real. Hennig y Kutlukaya (2007, p.21) abordan esta zona gris en su artículo recomendando lo que caracteriza a las funciones de pérdida en general, a saber, decisiones informales ante criterios de verdad relativos,

> „The main message of the paper is that the task of choosing a loss function is about the translation of an informal aim or interest that a researcher may have in the given application into the formal language of mathematics. The choice of a loss function cannot be formalized as a solution of a mathematical decision problem in itself, because such a decision problem would require the specification of another loss function. Therefore, the choice of a loss function requires informal decisions, which necessarily have to be subjective, or at least contain subjective elements."

6.8.1.4 Actualidad de los factores de Bayes: la elección de la Prior

Se dice de Jeffreys (1939/1961) que se interesó principalmente por las pruebas de significancia, lo que también se manifestó en su interés por los factores de Bayes. Al mismo tiempo, rechazó la prueba de significancia clásica de Fisher basada en valores *p* por muchas razones. A pesar de los comentarios críticos desde el punto de vista bayesiano (de Andrew Gelman o John K. Kruschke, entre otros), los factores de Bayes están gozando de sólo una creciente popularidad general (por ejemplo, Goodman, 1999; Held & Ott, 2018). Otros autores, como Gelman (2009b) y Gelman y Rubin (1995), rechazan fundamentalmente los factores de Bayes.

Las críticas a los factores de Bayes provienen de una visión clásica del enfoque Neyman-Pearson de Schimmack (2015c, 2016d) basada en diversos estudios, entre ellos el de Bem (2011a). Uno de los enfoques se centra en la distribución a priori, que el autor discute en detalle utilizando cálculos de ejemplo. Por lo tanto, continuamos la discusión del tema iniciada en el capítulo 4.4.2.2 con un enfoque ligeramente diferente, centrándonos ahora en la *elección de la Prior* en el contexto de los factores de Bayes (Liu & Aitken, 2008). La actualidad de la discusión puede verse en la discusión del estudio de Bem (2011a) sobre la clarividencia, que se cita a menudo. Para poder criticar el estudio de Bem, Wagenmakers, Wetzels, Borsboom y van derMaas (2011) recurren a factores de Bayes.

Schimmack (2018e) señala en otra entrada del blog, siguiendo un título de un artículo de Morey, Hoekstra, Rouder, Lee y Wagenmakers (2016), estos autores "quieren que los psicólogos cambien la forma en que analizan sus datos." Según Schimmack, la acusación general es que casi todo el mundo malinterpreta y entiende mal el trabajo clásico de Neyman-Pearson sobre los intervalos de confianza. Y este malentendido es la causa de la actual crisis de replicación en las ciencias sociales y en la psicología, respectivamente. Sin embargo, según Schimmack, a los autores no les preocupa la crisis de replicación per se, sino que "[l]a razón es que quieren utilizar la crisis de replicación como vehículo para vender la estadística bayesiana". Esta es

un punto de vista interesante, porque desplaza la discusión de la estadística pura a una dimensión política y social, cuya conveniencia no podemos ni queremos responder aquí, pero nos recuerda en este punto el proceso social de cambio de paradigma descrito por Kuhn (1973). Tales situaciones nos indican que los temas controvertidos a menudo no están necesariamente arraigadas en el contenido del tema, sino que están muy impulsadas por el propio interés individual – y esto presumiblemente se aplica a todos los implicados -- en representar y difundir la propia opinión con mayor o menor éxito y de forma lobbista. Nada más ya lo practicaba Ronald A. Fisher. En varias entradas de blog, Schimmack (2018e, 2015c) hace referencia al trabajo original de Neyman-Pearson (Neyman, 1937) para explicar la del factor de Bayes. Al hacerlo, él (Schimmack, 2018e) llega a la conclusión,

> „It is sad and ironic that Wagenmakers' efforts to convert psychologists into Bayesian statisticians is similar to Bem's (2011) attempt to convert psychologists into believers in parapsychology; or at least in parapsychology as a respectable science. [...] The problem with Bem's article is not the way he 'analyzed' the data. The problem is that Bem violated basic principles of science that are required to draw valid statistical inferences from data. It would be a miracle if Bayesian methods that assume unbiased data could correct for data falsification. [...] So, the problem is not the use of p-values, confidence intervals, or Bayesian statistics. The problem is abuse of statistical methods."

El trabajo de Morey, Hoekstra, Rouder, Lee y Wagenmakers (2016) citado anteriormente también es analizado por Gelman (2014c) – bajo el aspecto de la elección de la Prior, que no es trivial en absoluto y debería basarse en información contextual. Gelman (2014a) cuestiona si tiene sentido sustituir los valores *p* por intervalos de confianza de forma generalizada (véase también Gelman, 2013c) y concluye que los intervalos de confianza están sobreestimados. Si los factores de Bayes se ocupan principalmente del aspecto de la comprobación mediante métodos bayesianos, las suposiciones previas (las Priores) que entran en las ecuaciones son obviamente de importancia central. En el curso de la investigación psicológica (efectos pequeños, elevada proporción de ruido en los datos y las estimaciones), según el autor, las hipótesis a priori uniformes resultan a menudo problemáticas y tienen un efecto distorsionador, ya que se asigna a los intervalos de datos una significación sin que ésta esté justificada en términos de contenido. Esto ilustra que se debe ser capaz de cambiar de forma flexible entre la estadística clásica y la bayesiana para poder investigar el tema de forma exhaustiva (véase más arriba el capítulo 6.8.1.6 para un ejemplo de datos en R).

Otro punto de crítica se refiere a la aparición de los factores de Bayes. Para la crítica de Morey, Hoekstra, Rouder, Lee y Wagenmakers (2016), los factores de Bayes se calcularon a partir de valores *t*, grados de libertad e información adicional a priori y representan una *prueba t bayesiana general* (Rouder, Speckman, Sun,Morey & Iverson, 2009). Esto significa sin embargo, que todos los problemas que se aplican a los valores *p* ahora también son válidos para los factores de Bayes (Schimmack, 2015b), ya que ambas cantidades pueden derivarse de los valores *t* y los grados de libertad. Sin embargo, la interpretación de los resultados sigue siendo diferente. Mientras que los valores *p* adquieren un significado constante independiente del tamaño de la muestra, ya que asumen condiciones asintóticas de infinito, los factores de Bayes para el mismo valor *t* cambian cuando varía el tamaño de la muestra (Schimmack, 2016d, Kass, 1993). En concreto, esto significa que para los factores de Bayes, un mismo efecto en la población puede producir resultados diferentes dependiendo de lo grande que sea la muestra (véase la Tabla 6.10). Una simulación lo demuestra (véase la Fig. 6.18). La función de R `sim.p.bf()` genera muestras aleatorias para valores iniciales definidos para dos poblaciones y, a continuación, aumenta la diferencia media entre las dos poblaciones en las siguientes extracciones aleatorias. Los cuatro gráficos muestran la evolución de los factores de Bayes (BF_{01} en rojo, BF_{10} en azul) frente al valor *p* asociado (prueba *t* de Welch). Las líneas verticales señalan límites comunes como $p_{crit} = 0.05$ para el valor *p* (verde), así como $BF = 1$ ($= BF_{01} \approx BF_{10}$, naranja) y $BF = 3$ (= efecto significativo, rojo oscuro, véase Jeffreys, 1939/1961). Por gráfico las simulaciones `nsim` se realizan con el mismo valor inicial del generador aleatorio, que se regula mediante la variable `usesameseed` (`ptII_quan_Bayes_BayesFactors_dependence-on-N-sim.r`).

```
# simulate BF01 for various N R-Code
mu1 <- 100
```

```
mu2 <- 100
sigma1 <- 10
sigma2 <- 10
# all in one as a loop
# growing sample size, same sample size
par(oma=c(2,1,4,1), "cex.axis"=1, bty="l", mfrow=c(2,2))
ens <- c(30,90,250,1000)
for(n in ens) res <- sim.p.bf(N1=n,N2=n,mu1,mu2,
       sigma1,sigma2,nsim=1000)
mtext("Simulation Bayes Factors", outer=TRUE, line=2, cex=1.5, side=3)
mtext(expression(paste("dependence of ",BF[0][1]," on N",sep="")),
       outer=TRUE, line=0, cex=1, side=3)
```

Los gráficos de la Figura 6.18 muestran de forma impresionante la relación entre los valores *p* y los factores bayesianos en función del tamaño de la muestra (véase también Schimmack, 2015b). Según Schimmack (ibíd.) las consecuencias (véase la Tabla 6.10), a saber

> „This means that the same population e ect size can produce three di erent outcomes depending on sample size; it may showevidence in favor of the null-hypothesis with a small sample size, it may show inconclusive results with a moderate sample size, and it may show evidence for the alternative hypothesis with a large sample size.“

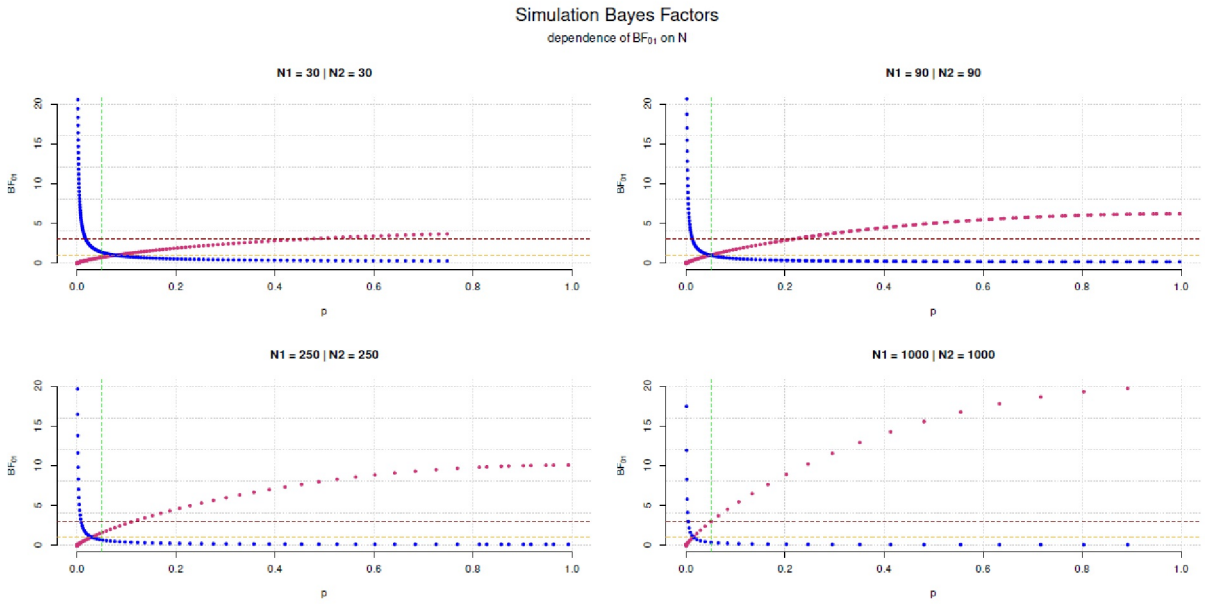


Figura 6.18: *Factor de bayes y tamaño de muestra*

Si la muestra sigue siendo del mismo tamaño, los factores de Bayes pueden calcularse directamente a partir de los valores *p* más Priores y viceversa. Sellke, Bayarri y Berger (2001) ofrecen en su artículo fórmulas para calibrar los *p*-valores con el fin de generar factores de Bayes. Los autores de `BayesFactor` ofrecen `ttest.tstat()` para derivar factores de Bayes directamente de los valores *t*. Aquí sigue un ejemplo (`ptII_quan_Bayes_BayesFactors_dependence-on-N-sim.r`) a partir de los datos de Gürtler ya presentados anteriormente por Gürtler (2005). Investigamos si existen diferencias de edad con respecto al género.

```
# age differences between sexes f vs. m
bwplot(age ~ sex, data=diss.red)
tres.sex <- t.test(age ~ sex, data=dats, var.equal=FALSE)
tv.sex <- tres.sex$statistic
pv.sex <- tres.sex$p.value
# bf10 = in favor of the alternative hypothesis
Ns.sex <- table(dats$sex)
bf10.sex <- ttest.tstat(t=tv.sex, n1=Ns.sex[1], n2=Ns.sex[2])
```

Tabla 6.10: *Dependencia del factor de Bayes del tamaño de la muestra*

Tamaño de la muestra	*Decisión*
pequeño	pro H_1
medio	indiferente entre H_1 y H_2
grande	pro H_2

Lo siguiente proporciona la salida:

```
> # output
> # Cohen's d
> cohensd(age[sex=="m"],age[sex=="w"])
d|mean sd d|pooled sd
0.14340 0.14446
> # t-value and p-value
> tv.sex
t
-1.2546
> pv.sex
[1] 0.2107
> # BF10
> exp(bf10.sex[['bf']])
[1] 0.26506
> # BF01
> 1/exp(bf10.sex[['bf']])
[1] 3.7727
```

La Prior ya está fijada por los autores y sólo se puede cambiar el factor de escala `rscale`. Aitkin (1997) ofrece más calibraciones de los valores *p*. La calibración de Sellke, Bayarri y Berger (2001) es relativamente sencilla de aplicar. Si se cumple p < 1/e, entonces el límite inferior del factor de Bayes (ibíd., p.62, fórmula 2) es

$$\text{minBF}_{01}(p) = \begin{cases} -e \cdot p \cdot \log(p) & \text{para } p < \frac{1}{e} \\ 1 & \text{si no} \end{cases} \tag{6.54}$$

y para la tasa de error frecuentista de tipo I para rechazar la H_0, se utiliza la siguiente fórmula (ibíd., p.62, fórmula 3).

$$\alpha(p) = \frac{1}{1 + \frac{1}{-e \cdot p \cdot \log(p)}} \tag{6.55}$$

Si el factor de Bayes creado de este modo se multiplica por las Priori Odds de H_0 o H_1, se obtienen las finales Odds de la Posterior. La primera calibración corresponde, según los autores, a una probabilidad posterior bajo el supuesto de que H_0 y H_1 tienen las mismas probabilidades a priori de $p = 0.5$, lo que debe justificarse en la práctica. La implementación en R es sencilla. Demuestra lo fácil que es implementar en R las fórmulas de los artículos. Reproducimos con la función de R `BF.calib()` los de Sellke, Bayarri y Berger (2001, p.63, Tabla 1).

```
# minimal BF / lower bound
# Sellke et al. 2001
# p-> # p-values from table 1 p.63
> pv <- c(.2,.1,0.05,.01,0.005,0.001)
> res <- sapply(pv, function(i) BF.calib(pv=i))
> colnames(res) <- pv
> print(res, digits=3)
         0.2   0.1   0.05  0.01  0.005  0.001
BF_p    0.875 0.626 0.407 0.125 0.0720 0.0188
alpha_p 0.467 0.385 0.289 0.111 0.0672 0.0184
```

Esta función R puede encontrarse en el paquete R `evidence` como `p2BF()`. Más recientemente, se ha puesto a disposición un paquete R llamado `pCalibrate` que permite derivar el factor de Bayes mínimo dentro de una clase especificada de hipótesis a partir de un valor *t* o *p* (Held & Ott, 2018). Por lo tanto, se trata de nuevo de una transformación de datos y no de una implementación bayesiana completa. El paquete de R cubre pruebas comunes como la prueba *z-*, *t-*, *F-* y la prueba de Ratio Likelihood. Basándose en el trabajo de Sellke, Bayarri y Berger (2001) mencionado anteriormente, Held y Ott (2016) tratan la dependencia de los factores de Bayes del tamaño de la muestra y determinan valores límite cuando este se aproxima a infinito. Además, amplían la calibración del valor *p* mediante otra fórmula (Held & Ott, 2018); y la comparan con variantes ya existentes:

$$\text{minBF}_{01}(p) = \begin{cases} -e\,(1-p)\,log(1-p) & \text{para } p < 1 - \frac{1}{e} \\ 1 & \text{si no} \end{cases} \qquad (6.56)$$

La comparación de los factores de Bayes generados por calibración se hace más interesante en el artículo de Held y Ott (2018, Tab. 3). La figura 6.19 visualiza los valores. Aquí, se calculan en función de la fórmula de calibración para diferentes p-valores comunes los factores de Bayes.

(`ptII_quan_Bayes_BayesFactors_dependence-on-N-sim.r`)

```
# comparison of minBF(p) approaches
# Held & Ott 2018 p = 0.05 (table 3)
minbfs <- matrix(data= c(1/2.1,1/2.5,1/3.4,1/3.9,1/6.8,1/7.5,
           1/6.5,1/8,1/14,1/15,1/28,1/37,
           1/11,1/14,1/26,1/28,1/51,1/74,
           1/41,1/53,1/112,1/118,1/224,1/368),
           ncol=4, byrow=FALSE)
colnames(minbfs) <- c("0.05","0.01","0.005","0.001")
minbfs
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l")
plot(minbfs[,1], ylim=c(0,1), type="b", bty="n", pch=20, xaxt="n",
           col="violetred3", pre.plot=grid(),
           xlab="Formula/ equation used to calculate BF",
           ylab="minBF(p)", main="")
for(i in 2:4) lines(minbfs[,i], col=i, type="b", pch=20)
axis(1, 1:6, paste("eq.",1:6,sep=""))
legend("top", legend=colnames(minbfs), col=c("violetred3",2:4),
           title="p value", bty="o", lty=1, lwd=2, cex=0.9,
           box.col="white", horiz=TRUE, bg="gray90")
mtext(expression(paste("Comparison of minBF(p) approaches",sep="")),
```

```
          outer=TRUE, line=-2.5, cex=1.5, side=3)
apply(minbfs, 2, summary)
apply(minbfs, 2, sd)
apply(minbfs, 2, fivenum2)
```

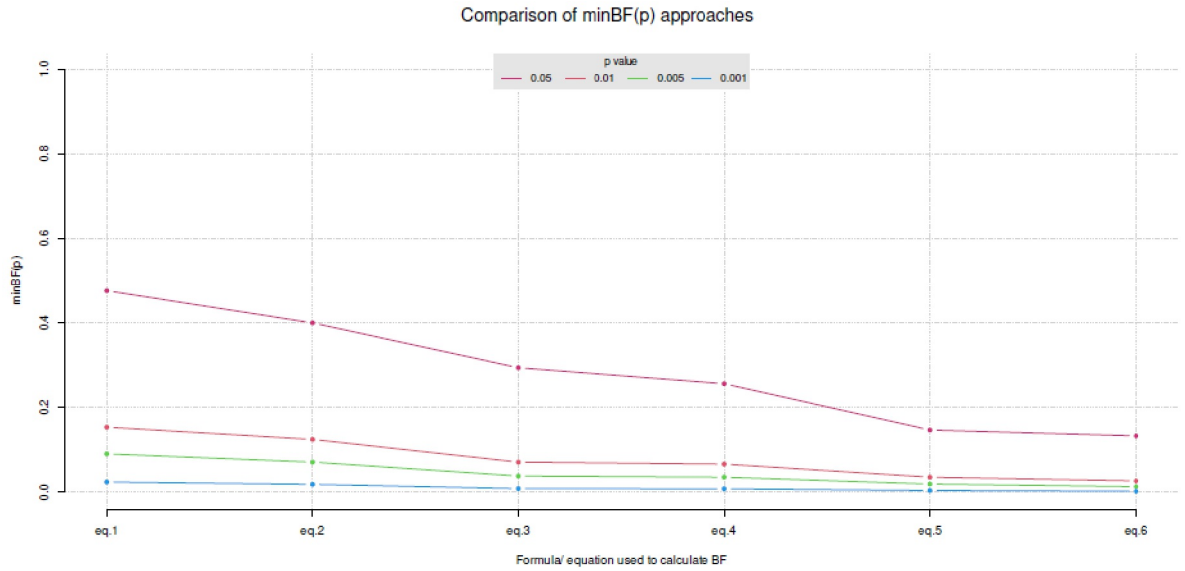


Figura 6.19. *Held y Ott (2018, factor de Bayes mínimo).*

Si a continuación se forma un cociente dividiendo el valor máximo por el valor mínimo de *minBF*, se obtiene

```
> # ratio range max and minBF
> apply(minbfs, 2, function(i) i[1]/i[6])
  0.05   0.01  0.005  0.001
3.5714 5.6923 6.7273 8.9756
> 1/(apply(minbfs, 2, function(i) i[1]/i[6]))
  0.05   0.01   0.005    0.001
0.28000 0.17568 0.14865 0.11141
```

Se observan ratios de entre 3.57 y 8.98 en el sentido de que estas diferencias son tanto mayores cuanto menores son los *p*-valores. Sin embargo, tales diferencias son ya de por sí tan grandes que se pregunta dónde está aquí la coherencia de la conclusión, que en realidad debería caracterizar al bayesiano (Cox, 1961; Jaynes, 2003). Si, en la práctica de la investigación, se elige entonces la ecuación que le proporcione el factor bayesiano mayor o menor o la media mínima? Tal incertidumbre dista mucho de ser suficiente para justificar un procedimiento de análisis de datos. No en vano, los autores profundizan en las respectivas ventajas e inconvenientes de las ecuaciones utilizadas, que se consideran relativamente independientes del tamaño de la muestra dentro de ciertos límites debidos a la calibración, siempre que se cumplan las condiciones límite elaboradas en el artículo (Held & Ott, 2018). Sin embargo, no son en absoluto independientes del tamaño de la muestra, como puede verse en los distintos gráficos y análisis del artículo. Teniendo en cuenta los valores *p* como base, esto no es sorprendente (véase la Fig. 6.20).

```
# simulate depence of p-values and N (sample size)
pes <- pvsN.sim()
pes
log(pes)
```

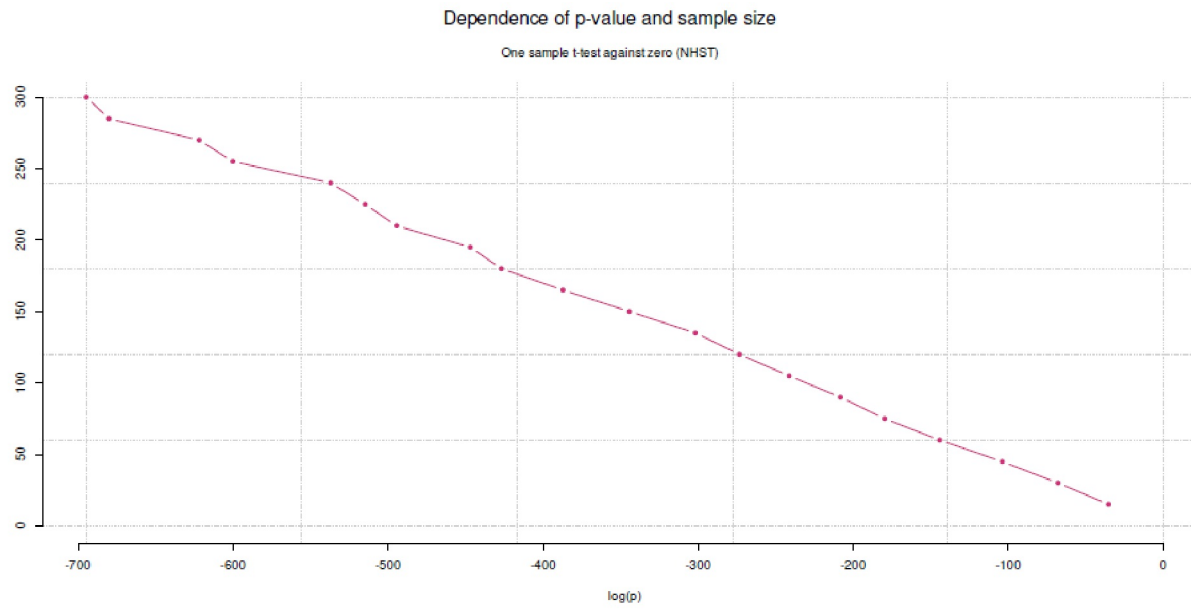


Figura 6.20: *Dependencia de valores p y tamaño de muestra*

Además, merece la pena examinar *mediante análisis de texto* el término calibración (algo abreviado) en los planteamientos aquí presentados. El significado del término *calibración* procede de la metrología y describe un proceso de documentación y determinación de la desviación de un instrumento de medida respecto a una referencia de forma válida y repetible. Un paso posterior tiene en cuenta la desviación de medición encontrada para futuras mediciones con este dispositivo de medición con el fin de mejorar la precisión de la medición y compensar las desviaciones. Sin embargo, la calibración no suele implicar un cambio del instrumento de medida como tal. Transferido a la situación de calibrar los valores *p*, esto sugiere que no se cambian sino que se describen sus desviaciones respecto a una referencia verdadera y se compensan a modo de prueba. En esencia, los factores de Bayes, que se determinan mediante la calibración de los valores *p* o sus bases (valores t, grados de libertad), siempre hay un valor *p* y, por tanto, todos los problemas de los *p-valores* (véase también el capítulo 4.3.9.1) están implícitamente presentes, aunque se intente determinar y compensar estos problemas (desviaciones). Debe considerarse en cada caso individual hasta qué punto esto realmente tiene éxito. No es de extrañar que las calibraciones de valores *p* sean especialmente defendidas por los representantes de la dirección bayesiana objetiva (por ejemplo, Sellke, Bayarri & Berger, 2001), cuyo principal objetivo parece ser el control total de los datos del análisis y el intento simultáneo de limitar fuertemente la influencia de la *supuesta* subjetiva Prior (véase también el capítulo 6.14.6). Esto recuerda al pensamiento frecuentista. Como puede verse en las ecuaciones anteriores, tiene lugar una transformación de datos no lineal, en principio reversible, de los valores *p*. En términos de análisis de textos, esto significa que se adapta algo y no que se crea algo nuevo. Se trata de una transformación sin cambiar el núcleo. Un cambio del núcleo requeriría, por ejemplo, una elección de una Prior adecuada al caso, así como el trabajo con la Posterior y, por tanto, la aplicación completa del teorema de Bayes. Ambas cosas sólo tienen lugar de forma limitada, si es que tienen lugar, en la calibración del valor *p*.

Si se considera – de modo texto analítico – lo que implica un proceso totalmente objetivo basado en datos es ante todo la ausencia de un marco de referencia y de contexto. Los datos, sin embargo, como un criterio relativo de verdad, no pueden hablar por sí mismos, aunque esto sea exactamente lo que puede leerse en algunos artículos profesionales y bien puede tener una validez pragmática subjetivamente experimentada por el individuo. Sin embargo, esta experiencia queda entonces limitada al sujeto de la investigación. En el frecuentismo, por tanto, uno se esfuerza por garantizar que el análisis sea válido en condiciones de infinitud con la esperanza de que las justificaciones dependientes del contexto queden entonces obsoletas. Las desviaciones se consideran errores (de medición, etc.) del valor verdadero que se quiere medir. Que esto puede fallar inevitablemente en un contexto estrecho y, además, conducir a afirmaciones que no

corresponden a las preguntas (sobre todo *p(D|H) vs. p(H|D)*) lo justifica detalladamente Jaynes (2003). En bayesiano, este problema básico se intenta resolver haciendo que la Prior, como estado actual de la falacia, establezca un punto de referencia respecto al cual se consideran los datos. La Posterior, como combinación o integración de la Prior y los datos (Likelihood), actualiza entonces este punto de referencia. Este proceso puede repetirse a voluntad, mientras que en la estadística clásica el final se alcanza tras la determinación de infinitas condiciones: ¿dónde debemos ir a partir de ahí? La única manera es pasar a la siguiente muestra aleatoria. En la bayesiana, el nivel relativo de significado de la perspectiva o punto de referencia y de los datos se mantiene en principio, sin derivar hacia condiciones imaginarias infinitas de las que sólo se pueden extraer muestras aleatorias, aunque pueden describir numéricamente una buena aproximación en la práctica de la investigación. Se trata de la mentalidad investigadora y la comprensión de los datos y el análisis, no de los números y las matemáticas. Sostenemos que si la atención se centra puramente en los datos que impulsan el análisis, debe establecerse algún tipo de punto de referencia. Si, como en el caso de los valores *p*, éste es infinito (es decir, asintótica), es decir, del cumplimiento de unas expectativas mínimas para poder calcular en absoluto (para que los supuestos de distribución puedan ser válidos, por ejemplo) surge la ilusión de que se puede prescindir de un punto de referencia. Lo que queda son probabilidades que sólo nos dicen algo sobre los datos, pero no sobre modelos, constructos e hipótesis. Si, por el contrario, se crea un punto de referencia, como en la estadística bayesiana, hay que vivir con el hecho de que es de naturaleza relativa. La relatividad como tal es irresoluble en el mundo experimentable sensorialmente. Esto sólo deja el discurso intersubjetivo para permitir la diversidad de perspectivas y abarcar, discutir y cuestionar, posiblemente excluir, integrar, etc. tantos puntos de vista como sea posible.

Sin embargo, si los factores de Bayes se operacionalizan no a través de la calibración del valor *p*, sino a través del cambio en la expectativa (Priores) dados los datos, esto va de la mano con la elección de un valor a priori y, por lo tanto, va más allá de los valores *p*. Una vez más, ¿por qué calibrar los valores *p* cuando se dispone de un enfoque bayesiano completo a partir de los datos brutos? Un uso legítimo de las calibraciones de los valores *p* parece ser cuando, como en los metaanálisis, no se puede acceder a los datos brutos originales. Sin embargo, esto no nos exime del laborioso trabajo de determinar las Priores adecuadas para establecer el punto de referencia bayesiano, sin seleccionar globalmente las Priores en función de las características matemáticas. De lo contrario, no se trata de estadística bayesiana. Queda por ver si la elección de una Prior es posible en un metaanálisis sin más información, o si es posible en absoluto. La salida a través de Priores objetivas automáticas y gobernadas por reglas no parece ser una solución ideal, aunque hay representantes de la estadística de Bayes que piden precisamente esto. Se podría considerar si el término "Bayes" en el factor Bayes parece apropiado en absoluto en el contexto de una calibración de valores *p*. Ahora parece mucho más sensato utilizar simplemente el término *calibración del valor p* para esta clase de análisis, porque describe con bastante precisión lo que ocurre.

Ly, Verhagen y Wagenmakers (2016) enumeran otras diferencias entre los factores de Bayes y los *p-valores*. Un argumento significativo es que los valores *p* no se basan únicamente en los datos observados, ya que hacen afirmaciones sobre datos no observados más extremos que ellos mismos. Los factores de Bayes se refieren estricta y únicamente a los datos observados. Los datos (más) extremos que los valores *p* cubren por definición son irrelevantes; o como dice Wagenmakers (2007c, p.1), haciéndose eco de la famosa cita de Jeffreys, "Los procedimientos frecuentistas dependen de acciones hipotéticas para sucesos imaginarios". Mientras que en el ritual nulo (véase el capítulo 4.3.8) el valor *p* se basa en un único modelo frente a la clásica H_0, el factor de Bayes prueba dos modelos entre sí. Esto tiene lugar en un nivel relativo pero continuo – qué modelo está mejor respaldado por los datos observados. La *prueba de H_2 frente a H_1* con los factores de Bayes es estructuralmente similar a la *lógica de decisión* según Neyman-Pearson (véase el capítulo 77), pero no a la teoría de Fisher, que sólo se centra en H_0 (véase el capítulo 4.3.2) o incluso a la NHST (véase el capítulo 4.3.8). Schimmack (2015b) proporciona una discusión detallada de las implicaciones resultantes. Cabe mencionar que el factor de escala *r* de una Prior puede favorecer o no a la H_0 clásica. Si el factor de escala *r* – y no se dispone de criterios definitivos al respecto – se elige de tal forma que la Prior se hace más amplia y se concentra en el cero, la hipótesis asociada debe tener valores más extremos y se rechaza más fácilmente en el caso de efectos pequeños. Esto demuestra por sí solo lo cuidadosa y sustancial que debe ser la elección de la Prior o que, como es habitual, no existen criterios definidos sobre *cómo* debe elegirse la

Prior. Wagenmakers, Wetzels, Borsboom y van der Mass (2011), por ejemplo, utilizan una distribución Cauchy con factor de escala = 1. El paquete `BayesFactor` de R establece `medium` (rscale = 1/2) como valor predeterminado. Además, existen las opciones `wide` (rscale = $\sqrt{2}/2$) y `ultrawide` (rscale = 1).

Schimmack (2015b, 2015c) proporciona una discusión en profundidad con ejemplos numéricos e implicaciones de las dependencias de los factores de Bayes, los tamaños de muestra, los valores *p* y los tamaños de efecto, así como la influencia de la Prior. La Figura 6.21 reproduce los resultados de Wagenmakers, Wetzels, Borsboom y van der Mass (2011) para diferentes factores de escala *r* en función del tamaño de efecto δ. Dado que la distribución se centra por encima de cero, se favorece esta hipótesis como más probable que otras. Aquí es donde se encuentra la masa de la distribución y, por lo tanto, la ponderación con respecto a H_0. Un factor de escala *r = 1* significa, por ejemplo, que el 50% de la masa de la distribución se sitúa entre -1 y +1. El equivalente se aplica a los demás factores de escala *r*. Los tamaños del efecto por encima y por debajo del valor absoluto del factor de escala *r* se ponderan por igual. Cuanto mayor sea este factor, menos probable es que se favorezca la hipótesis nula porque la distribución se hace más amplia. En la versión con un factor de escala *r* de `rscale` = $\sqrt{2}/2$, sin embargo, (Schimmack, 2015b) una parte considerable de la ponderación sigue recayendo en tamaños del efecto superiores a d = 2. Como recordatorio - según Cohen (1969) los tamaños del efecto de d > 0:8 ya se consideran efectos fuertes y un d > 2 debe alcanzarse primero de forma experimental o no experimental. Además, como variable descontextualizada, un tamaño del efecto debe, no obstante, estar anclado en el contexto, es decir, ¿qué tamaños del efecto se utilizan habitualmente en esta o aquella situación de investigación en condiciones que se especificarán con más detalle? Así pues, no está claro por qué una Prior considera que tales valores son relativamente probables. La elección de Prior tiene consecuencias. Por ejemplo, el factor de Bayes puede favorecer la hipótesis nula porque la población tiene un efecto pequeño verdadero, pero la Prior considera que incluso los efectos extremadamente grandes son muy posibles y, por tanto, los efectos pequeños no se corresponden con las expectativas a priori. Los factores de Bayes contienen (Schimmack, 2015b) la información "de que los factores de Bayes expresan la fuerza relativa de las pruebas a favor o en contra de la hipótesis nula en relación con el tamaño del efecto implícito por la función predeterminada". La relación entre el tamaño del efecto verdadero y el factor de escala *r* tiene este aspecto (`ptII_quan_Bayes_Bayes-Factors_dependence-on-N-sim.r`). Las hipótesis se denominan H_0 y (alternativamente) H_1 en el sentido clásico. También podríamos referirnos a ellas como H_1 y (alternativamente) H_2. En la práctica, no hay diferencia. Lo único que hay que saber es qué llamar en sentido estricto y cómo para poder sacar conclusiones posteriores correctamente.

- Si la intensidad del efecto real es mayor que el factor de escala *r*, el factor de Bayes favorece a H_1. A medida que disminuye el error de muestreo, el factor de Bayes crece hacia 1.
- Si la intensidad del efecto real es menor que el factor de escala *r,* el factor de Bayes favorece inicialmente a la H_0, ya que la H_1 contiene (demasiadas) intensidades de efecto que incluyen valores más extremos. A medida que aumente el tamaño de la muestra, el factor de Bayes favorecerá a H_1.

La Figura 6.21 muestra esta distribución de Cauchy con el factor de escala rscale = $\sqrt{2}/2$:

```
# plot cauchy for ES R-Code
# see how Wagenmakers et al. (2011) work
# and critized by Schimmack (2015) on his blog
sek <- seq(-4,4,0.01)
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l")
plot(sek, dcauchy(sek,scale=sqrt(2)/2), type="l", bty="n",
       col="darkred", pre.plot=grid(),
       xlab=expression(theta),
       ylab=expression(paste("density with r=",sqrt(2)/2,sep="")),
       cex.lab=1.2)
mtext(expression(paste("Cauchy Distribution with scale = ",
       sqrt(2)/2,sep="")),
       outer=TRUE, line=-3, cex=1.5, side=3)
```

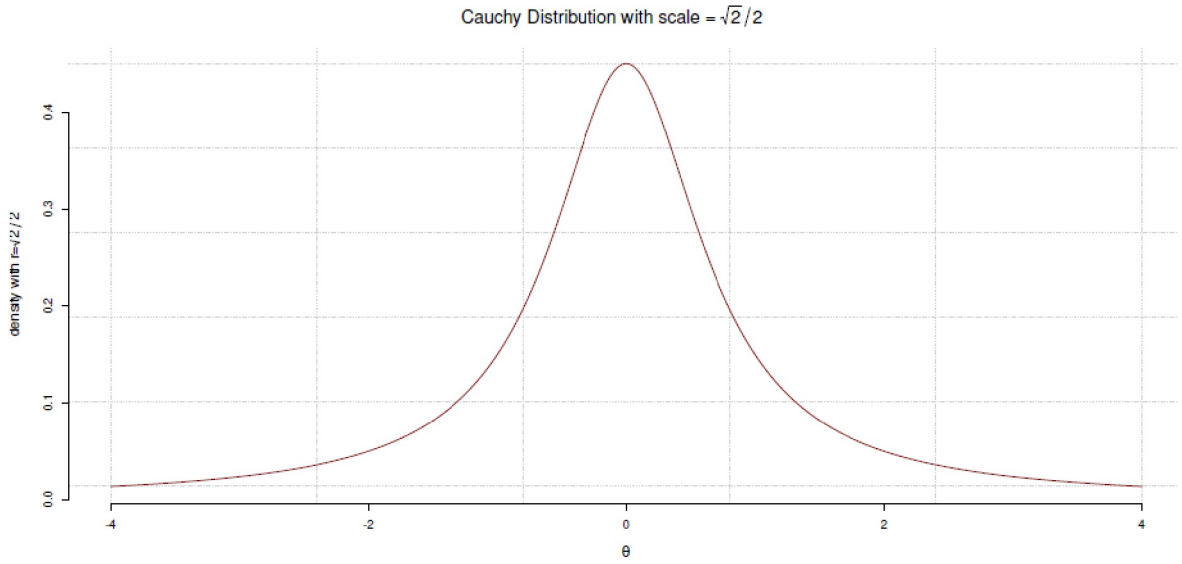


Figura 6.21: *Distribución Cauchy con factor de escala*

Después Schimmack discute con ejemplos númericos la dependencia de la Prior y el tamaño de muestra en los resultados de Wagenmakers, Wetzels, Borsboom y van der Mass (2011) en cuanto al estudio de Bem. Su argumento principal es, que un factor de escala *r* de `rscale = 1` para la Prior (distribución Cauchy) es no realista en el ámbito de la investigación psicológica, ya que se observa raramente potencias de efecto $d > 1$. Además, dado una potencia de efecto $d > 1$ no se necesitaría más estadísticas particulares, pues un sencillo gráfico como un gráfico dispersión o un boxplot sería suficiente para rechazar la H_0 convincentemente. Por eso, Schimmack (2015c) concluye

> „The strength of evidence for an effect is predictable from the precision of the alternative hypothesis. [...] The strongest support for an effect is obtained for the uniform distribution with a range of effect sizes from .1 to .3. The advantage of this range is that the lower bound is not 0. Thus, effect sizes below the lower bound provide evidence for H0 and effect sizes above the lower bound provide evidence for an effect. The lower bound can be set by a meaningful consideration of what effect sizes might be theoretically or practically so small that they would be rather uninteresting even if they are real."

Además, el análisis de potencia según Neyman-Pearson, en el que entran los tamaños del efecto según Cohen (1969), espera un valor único. Esto no tiene sentido en un marco bayesiano, ya que se trata de las distribuciones posteriores completas y no sólo del valor modal de la Posterior. En este sentido, los rangos de valores que diferencian entre distintos niveles de importancia parecen una mejor opción, también para expresar una Prior. Así, cuanto más clara, direccional e inequívocamente se exprese una expectativa como Prior, más claramente se encontrará un efecto, pero también podrá rechazarse. Esto muestra la diferencia entre exploración y confirmación. En el caso de la confirmación, las hipótesis son muy específicas. En resumen hay que tener en cuenta que los factores de Bayes dependen de la elección y precisión de la H1 (hipótesis alternativa). Sin embargo, un efecto no puede formularse vagamente y sin sentido, sino que requiere una especificación clara de la dirección y la magnitud (véase el capítulo 4.3.3.2) para que se corresponda con la expectativa. Esto presupone que una hipótesis sobre un efecto se formula con precisión en términos de contenido. Schimmack señala: "Resulta sorprendente, por tanto, la poca atención que los estadísticos bayesianos han dedicado a la cuestión de la especificación de la distribución a priori."

En el caso de la citada crítica al estudio de Bem por parte de Wagenmakers, Wetzels, Borsboom y van der Mass (2011) y Wagenmakers, Wetzels, Borsboom, Kievit y van der Maas (2011) respectivamente la Prior

es justificado por los autores en base a los desiderata de Jeffreys (1939/1961) y la preocupación por la objetividad, sin abordar los tamaños del efecto intersubjetivos bajo la H_1, es decir, aquellos que se justifican en el tema y en base a estudios previos y que, por tanto, son esperables. Se trata de una omisión, ya que sigue sin estar claro cómo se conecta este enfoque con el contexto del contenido. En consecuencia, se adopta un punto de vista conservador, pero no adecuado al caso. Para investigar la cuestión de la objetividad de la Prior para el tamaño del efecto bajo la H_1, a saber, $p(\delta \mid H_1)$, se realizó un análisis de robustez (Wagenmakers, Wetzels, Borsboom, Kievit & van der Maas, 2011), p.2f.),

> „We therefor carried out a robustness analysis in which we systematically varied the scale parameter for p(δ | H1), and reported the results in an online appendix3. These results showed that for a wide range of different, non-default prior distributions on effect size the evidence for precognition is either non-existent or negligible."

Schimmack (2015c) critica este procedimiento porque la Prior está centrada sobre cero y se realizó una prueba bilateral, lo que dificulta separar la H_0 de la H_1. Solo se cambiaron los parámetros de Wagenmakers, Wetzels, Borsboom y van der Mass (2011), pero ni la distribución de Cauchy ni el centrado por encima de cero. Según Schimmack (2015c), se podría haber encontrado un efecto para el estudio de Bem si se hubiera incluido una Prior clara y dirigida en el teorema de Bayes. Sin embargo, Schimmack no considera que el estudio de Bem sea un buen estudio empírico con resultados válidos ni mucho menos, ya que también podría ser que los resultados se obtuvieran mediante prácticas de investigación dudosas que introducen el error de tipo I (de forma inflacionaria (Francis, 2012; Schimmack, 2012; Schimmack, 2015c). Tal aclaración requiere la replicación por parte de otros investigadores, lo que aún no se ha conseguido (véase también el capítulo 4.4.2.2). Esto crea una situación en la que el problema (Gelman & Loken, 2013, 2014) no radica en la naturaleza de los análisis estadísticos en sentido estricto, sino más bien en que los datos no parecen ser suficientemente fiables (Schimmack, 2015c):

> „However, the problem is not that the data were analyzed with the wrong statistical method. The reason is that the data are not credible."

Un cambio de frecuentista a bayesiano de vuelta a frecuentista no cambia nada en absoluto en una situación tan crítica. Esto solo se puede cambiar mediante réplicas serias que tengan suficiente potencia en el sentido de Neyman-Pearson (Schimmack, 2016c). Basándose en varios estudios con datos fiables, se puede entonces trabajar meta-analíticamente. Según Schimmack (2015b), surgen problemas en el contexto de los factores de Bayes equivalentes a los del enfoque de Neyman-Pearson (véase el capítulo 4.3.3):

- Dependencia de la muestra (tipo de muestra o tamaño).
- Sólo evidencia relativa de la superioridad de una hipótesis sobre otra. ¿De dónde debe venir la absolutidad?
- Separación poco clara de cuándo las pruebas hablan claramente a favor o en contra de una hipótesis o de una zona gris indiferente (véase también la teoría de la decisión y los problemas fundamentales en la elección de umbrales críticos). Sin embargo, se trata de un problema básico de toda estadística, ya que afecta a las transiciones de parientes empíricos a parientes numéricos.
- Las funciones estándar pueden no representar adecuadamente la hipótesis alternativa o el modelo competidor (hipótesis). La Prior del tamaño del efecto influye en el factor de Bayes y, por tanto, en la preferencia de este por una de las hipótesis en competencia.

Schimmack (2015b) concluye la entrada de su blog con

> „In conclusion, like p-values, Bayes-Factors are not wrong. They are mathematically defined entities. However, when p-values or Bayes-Factors are used by empirical scientists to interpret their data, it is important that the numeric results are interpreted properly. False interpretation of Bayes-Factors is just as problematic as false interpretation of p-values."

También se puede resumir esto diciendo que el problema básico, que ninguna estadística puede resolver, a saber, la dependencia de los análisis a lo largo de los factores de influencia del contexto y la situación, los supuestos del modelo, el conocimiento previo, los parámetros, etc., sigue existiendo. Sólo se puede intentar estimar seriamente las consecuencias y comunicar claramente las zonas grises. Las funciones de pérdida pueden ayudar a ello. Hasta el más mínimo indicio de panaceas (semi)automáticas que siempre funcionan debería despertar sospechas masivas.

Dado que los factores de Bayes se utilizan a menudo en estudios con un menor número de casos, la elección de la Prior es esencial, ya que el factor de Bayes reacciona sensiblemente a los cambios en la Prior.

Al parecer, el propio Jeffreys (1939/1961) creó las Priores del factor de Bayes basándose en las funciones de Likelihood, ya que su objetivo era desarrollar una prueba con una amplia gama de aplicaciones. En este sentido, su trabajo sirve como punto de referencia que debe mejorarse o adaptarse situacionalmente mediante el conocimiento específico del contexto. Desde esta perspectiva, se pueden entender las Priores como funciones de ponderación. Un modelo de pensamiento ilustra esta forma de pensar: En el caso de que falte información previa, debe elegirse una ponderación tal que $BF_{21}(prior) = 1$, es decir, que ambas hipótesis sean igualmente probables. En el caso de una hipótesis infinitamente precisa que muestre un efecto, la ponderación debe elegirse de modo que $BF_{21}(prior) = \infty$ (Ly, Verhagen y Wagenmakers, 2016)). En el caso de modelos competidores anidados, es decir, los modelos son idénticos excepto por el número de parámetros en un modelo H_2 en comparación con H_1, se pueden realizar más ajustes. El uso de factores de Bayes para probar nuevos factores en modelos *anidados* es defendido por Kruschke (2011b), entre otros, aunque el autor critica por lo demás los factores de Bayes como criterio principal para detectar efectos. Kass (1993, p.555), sin embargo, señala que los factores de Bayes (pueden) ser sensibles a la Prior incluso con un gran número de casos. En un artículo de blog con código R reproducible, Schönbrodt (2013), por otra parte, examina la robustez de los factores de Bayes con respecto a diferentes valores *t* y tamaños de muestra. Como ejemplo, el autor reproduce el reanálisis de Wagenmakers, Wetzels, Borsboom, Kievit y van der Maas (2011; 2011b) sobre los datos del estudio de Bem. Examina los factores de Bayes (prueba unilateral según Morey y Wagenmakers, 2014) sobre la base de varias Priores y llega a la conclusión de que, al aplicar esta a los datos de Bem, todos los gráficos parecen ligeramente desplazados hacia H_2, pero no mucho.

Los defensores de los factores de Bayes argumentan que éstos, a diferencia de los *p*-valores, permiten recopilar datos hasta que los valores se estabilizan sin -- acumulación de tasa de error o el riesgo de socavar manipulativamente la propia investigación (Palabra clave: p-hacking, Rouder, 2014; Wagenmakers, 2007b para un resumen del estado del debate).

Si se busca explícitamente una prueba de hipótesis, los factores de Bayes permiten precisamente eso: probar hipótesis entre sí. La elección de la Prior – aunque los factores de Bayes son sensibles a ella – permite incorporar directamente el conocimiento previo, como es habitual en la estadística bayesiana, y examinar y aprender del comportamiento de los resultados en relación con este conocimiento previo. Nos preguntamos:

> Tarea 6.5: ¿Bicho o característica?
>
> ¿La sensibilidad a priori es ahora un error o una característica? ¿Por qué tantos investigadores tienen problemas con esto? La conexión es obvia, porque según el teorema de Bayes, el conocimiento previo y la probabilidad están relacionados multiplicativamente. Por tanto, la naturaleza de la dependencia de ambas cantidades debería ser obvia.

En combinación con una función de pérdida que se especificará con más detalle, se pueden investigar más a fondo las consecuencias respectivas de decidirse a favor de una u otra hipótesis o de considerar indiferente el resultado, de forma que se pueda obtener como resultado global una teoría de la decisión razonada.

La preocupación por los factores de Bayes y los artículos pertinentes plantea ahora también un problema bien conocido para la estadística de Bayes que ya causa problemas en la estadística clásica. La cuestión es: ¿en qué momento un factor de Bayes es no significativo, poco significativo, muy significativo o altamente significativo? Es lo mismo que preguntarse hasta qué punto es significativo un valor *p* o un tamaño del efecto. Independientemente del hecho que, por supuesto, los efectos difieren en su significación y las decisiones deben tomarse según criterios definidos, este debate no puede llevarse a cabo en abstracto y desvinculado del contexto. La proliferación de tablas, como la frecuente referencia al trabajo original de Jeffreys con reglas abstractas sobre cuándo y cómo clasificar una BF_{21} *como significativa*, hace precisamente eso. Descontextualiza los análisis y abre la puerta a barreras generales aprobadas por convención. El problema es ahora ampliamente discutido, ya sea a partir de caricaturas sobre los valores *p* (Munroe, s.f.), en forma de publicaciones y entradas de blog (Gelman, 2015ª; Kruschke, 2015c) o en artículos profesionales (Gigerenzer, 2004b; Gigerenzer & Marewski, 2015). El problema básico se reconoce en el hecho de que detrás de las reglas y barreras generales se esconde la idea de métodos y modelos analíticos automáticos universalmente aplicables y siempre válidos que funcionan en todas las circunstancias y lo mejor de todo por sí solos. Completamente pasado por alto es el hecho de que los métodos analíticos cambian retroactivamente la disciplina y lo hacen a largo plazo y por lo tanto tienen un impacto masivo en la conciencia y la percepción y el tratamiento de los objetos de investigación. Si surge la ilusión de que todo (por ejemplo, las conclusiones) es factible mediante algoritmos, la replicación, la minimización de los errores de medición y la mejora de los instrumentos y métodos de recogida de datos, así como otros factores, se descuidan en la planificación del diseño. Lo mismo sabemos de la neurociencia. Hace 20 años, bastaba con mostrar una imagen fMRI de un cerebro en una presentación y todos los presentes quedaban profundamente impresionados. Sin embargo, muchos (todavía) no saben que detrás de estas imágenes hay más de un 80% de ruido y enormes fallos metodológicos (Bennett, Baird, Miller & Wolford, 2009). Además, estos métodos de imagen no representan colores, sino únicamente valores en escala de grises, es decir, valores en una dimensión. Los colores surgen – como es de esperar – cuando se superan determinados umbrales predefinidos de actividad cerebral. Los problemas surgen, por ejemplo con los falsos positivos. Como recordatorio, los falsos positivos indican el falso rechazo de la hipótesis nula, es decir, la suposición de un efecto que en realidad no está presente. Estos problemas han llegado tan lejos que hay autores (Eklund, Andersson, Josephson, Johannesson, et al. Knutsson, 2012; Eklund, Nichols, y Knutsson, 2016b; Michael, Newman, Vuorre, Cumming, y Garry, 2013) que ponen en duda por razones de metodología y estadística (por ejemplo, debido a los fallos metodológicos en la identificación de voxels) gran parte de los estudios empíricos de neurociencia de las últimas décadas basados en imágenes (fMRI). Además de estas deficiencias también se critica la escasa potencia de los estudios, un hecho que la psicología conoce demasiado bien.

Sin embargo, si resumimos los argumentos de forma algo menos polémica, los factores de Bayes pueden considerarse otra fuente de información junto a las estimaciones de parámetros, las simulaciones y las evaluaciones gráficas (Gabry, Simpson, Vehtari, Betancourt & Gelman, 2019).Consideramos poco prudente basar las conclusiones en un único criterio de calidad. La paradoja de Lindley-Bartlett (1957, véase la sección 4.4.14.2) demuestra de forma impresionante que los factores de Bayes proporcionan pruebas de la validez (o no) de las hipótesis nulas y que este fenómeno puede contrastar claramente con los valores significativos *p*. Así pues, los factores de Bayes pertenecen a la categoría de *comprobación de hipótesis* y reúnen todas las ventajas y críticas que pueden asociarse a este tema, especialmente la problemática asignación de umbrales de cruce críticos y libres de contexto. Los factores de Bayes permiten aportar pruebas de la validez de la H_0 clásica, algo imposible en la estadística clásica. Allí, la H_0 sólo puede no rechazarse, lo que corresponde a una ganancia de conocimiento muy modesta (véase también el capítulo 4.4.9 sobre la cuestión de la equivalencia, que corresponde a una H_0 no rechazada). La prueba de la H_0 clásica puede ser científicamente muy interesante para anular el sesgo hacia una búsqueda rígida de los efectos existentes y desencadenar explícitamente preguntas que conduzcan al conocimiento cuando no se produzca ningún efecto. Toda investigación sobre efectos secundarios en el ámbito farmacéutico se basa en este modelo de pensamiento, en el que hasta ahora se han utilizado principalmente pruebas de equivalencia frecuentistas.

Así que ha llegado el momento de realizar un pequeño estudio sobre la clarividencia para la comprobación de hipótesis nulas - de modo *bayesiano*.

6.8.1.5 Ejemplo de investigación: ¿la clarividencia?

¿Es un estudio sobre la clarividencia un buen motivo para formular hipótesis nulas? La motivación de un pequeño estudio sobre la clarividencia no surgió tanto de las críticas al estudio de Behm (2011a) o de "probar" la clarividencia o incluso de lo contrario. Más bien, la idea era realizar un experimento junto con los participantes del seminario de forma manejable, en el que todos pudieran participar activamente. El tema de la clarividencia se presta excelentemente a tales fines. Especialmente para un experimento requiere muy poco trabajo previo. La variable dependiente era tradicionalmente la adivinación frente a la "clarividencia" de las cartas (corazones, diamantes, tréboles y picas). El orden de las cartas se generó aleatoriamente con R y `runif()`. Los sujetos fueron asignados aleatoriamente a dos grupos de la misma manera. Lo mismo se aplicó a la secuencia de sujetos para la serie de pruebas cortas. Para hacerlo más interesante, el grupo de tratamiento recibió ruido blanco en los auriculares durante el experimento y el grupo de control no. Además de adivinar las cartas, se preguntó a los sujetos cada vez sobre la certeza subjetiva con la que experimentaban la exactitud de la clarividencia de las cartas. Con N = 20 sujetos, hubo k = 20 repeticiones cada una, de modo que se crearon N*k = 400 datos, cada uno para la *clarividencia* y la *certeza subjetiva* respectivamente. Además, todos los participantes se calificaron a sí mismos en cuanto a si tenían una actitud positiva, neutra o negativa hacia la clarividencia en principio. Con una probabilidad de acierto de p = 1/4 = 0.25 por intento de acierto y con m = 400 ejecuciones, cabe esperar un porcentaje de aciertos típico de 400/4 = 100. Dado que también se añade el "ruido experimental" o como quiera que se llame la cantidad de influencias incontrolables, no debemos suponer exactamente 100, sino algo menos a algo más. El concepto de ROPE de Kruschke (2011b, 2015b, véase el capítulo 6.8.4.2) es precisamente adecuado para definir ese rango de tolerancia de efectos inexistentes. La prueba consiste en determinar si los sujetos se desvían de esta zona y no de la estimación puntual exacta de p = 0.25, ya sea hacia abajo o hacia arriba. Es necesaria una prueba bilateral porque una probabilidad de conjetura negativa también es muy improbable. Esto sería como la clarividencia, pero con signo invertido. La probabilidad binomial de acertar todas las k = 20 cartas con p = 0.25 es (`ptII_quan_Bayes_caso_exp-extra-percepción-sensual.r`).

```
> 0.25^20
[1] 9.0949e-13
```

9.09e -13 es significativamente inferior a ganar la lotería, con casi 1 entre 14 millones, respectivamente.

```
> 1/choose(49,6)
[1] 7.1511e-08
```

Por otra parte, la posibilidad de no obtener ni un solo acierto es muy pequeña, sólo

```
> # no hit
> .75^20
[1] 0.0031712
```

El análisis comienza con un vistazo a los datos, las tablas y algunos gráficos - véase también el capítulo 5 sobre el AED según Tukey (1977).

```
# read data
hell <- read.table("LG_clairvoyance-exp_raw_data.tab",
sep="\t", header=TRUE)
postsubj <- read.table("LG_clairvoyance-exp-post_raw_data.tab",
sep="\t", header=TRUE)
```

El análisis de la tabla de la variable dependiente diferente, es decir, si las tarjetas se reconocieron como VERDADERAS (= correctas) o FALSAS, apunta muy claramente hacia el principio aleatorio esperado, es decir,

la adivinación. Los gráficos tampoco indican que los sujetos de la muestra tuvieran capacidades perceptivas especiales. No obstante, a continuación lo examinaremos con más detalle para ver hasta qué punto puede cuantificarse la incertidumbre con respecto a la confirmación de la hipótesis nula.

La comparación de los grupos de tratamiento y control no indica diferencias significativas en la capacidad de percepción. Lo mismo muestra el diagrama de caja (véase la Fig. 6.22)

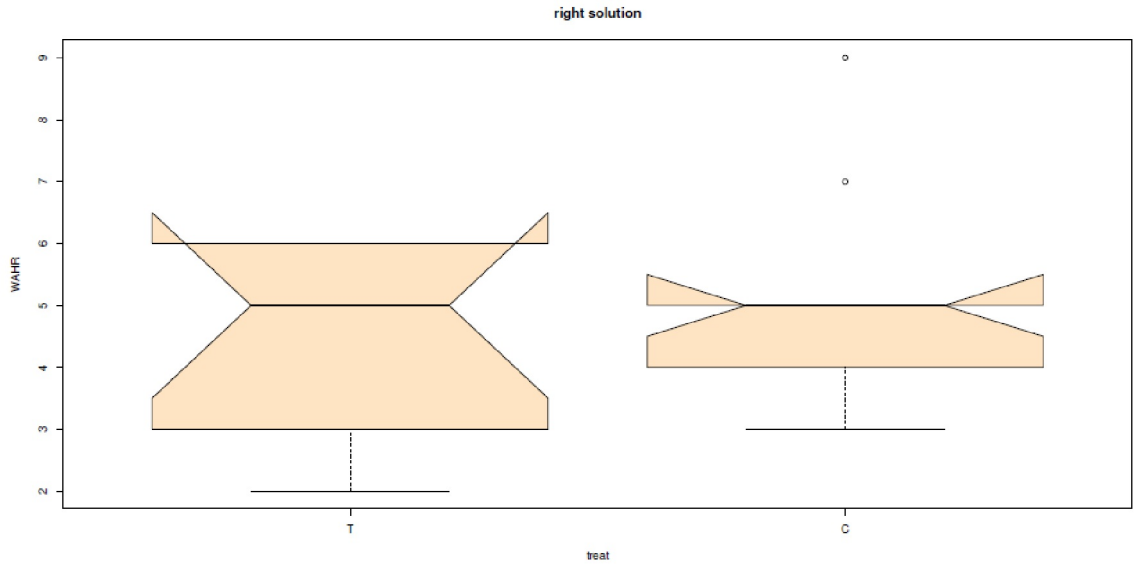


Fig. 6.22. *Estudio de la clarividencia*
(boxplots, grupo de tratamiento y grupo de control, soluciones correctas)

```
# prepare tables
head(hell)
CT <- table(hell$Upn,hell$treat)
CT.fac <- as.factor(CT[,"C"])
levels(CT.fac) <- c("T","C")
upn.differ <- with(hell, table(Upn, differ))
hell.res <- data.frame(cbind(Upn=dimnames(upn.differ)$Upn,
       upn.differ[,1:2]), treat=CT.fac)
hell.res$WAHR <- as.numeric(hell.res$WAHR)
hell.res$FALSCH <- as.numeric(hell.res$FALSCH)
hell$differ <- as.factor(hell$differ)
hell$differ.TF <- ifelse(differ == "WAHR", TRUE, FALSE)
hell$treat <- fas.factor(hell$treat)
# actual capability clarivoyance
with(hell.res, boxplot(WAHR ~ treat, col="bisque",
       notch=TRUE, main="right solution"))
```

La relación correlativa entre la certeza subjetiva y la capacidad de adivinación objetiva es de aproximadamente nada:

```
> # relationship correct solution and subjective confidence
> cor(as.numeric(hell$differ), hell$subjsicher)
[1] -0.014771
```

Un examen más detallado de los datos no indica que haya diferencias entre los grupos de tratamiento y de control en cuanto a la capacidad de clarividencia objetiva.

```
> # EDA
> tab.tf <- with(hell, table(solution, differ))
> tab.tf
differ
solution FALSCH WAHR
Corazón    74 29
Diamantes  71 23
Tréboles   75 19
Pica       85 24
> mosaicplot(tab.tf) # not shown
> apply(tab.tf,2,sum)
FALSCH WAHR
305    95
> 95/(305+95)
[1] 0.2375
> # summary
> upn.differ <- with(hell, table(Upn, differ))
> t(apply(upn.differ,2,summary))
       differ Min.  1st Qu. Median  Mean  3rd Qu. Max.
FALSO   11 14.75  15       15.25 16.25   18
TRUE     2  3.75   5        4.75  5.25     9
```

Un máximo de 9 de cada 20 cartas fueron reconocidas correctamente, y un mínimo de 2 de cada 20.

Desde un punto de vista estadístico clásico con `binom.test()`, esto no sería especialmente emocionante. Lo mismo ocurre con el enfoque bayesiano, realizado con `pbl()` y `pjc()`, los cálculos de probabilidad para tasas de éxito distribuidas binomialmente, que se presentan en más detalle en el capítulo 6.15.2 (Studer, 1996b). ¿Existen valores atípicos? es decir, ¿hay personas especialmente buenas o especialmente malas para predecir las cartas?

Dentro de la variabilidad esperada, estos individuos no existen. Por lo tanto, la probabilidad media esperada de acertar y el rango de los datos son suficientes

```
> # short summary statistics
> apply(upn.differ.bp,2,mean)
mean(pbl) mean(pjc)
0.26145   0.23750
> apply(upn.differ.bp,2,range)
    mean(pbl) mean(pjc)
[1,] 0.136 0.10
[2,] 0.455 0.45
```

con valores que oscilan entre 0.136 y 0.455. La figura 6.23 muestra los dos histogramas de las distribuciones.

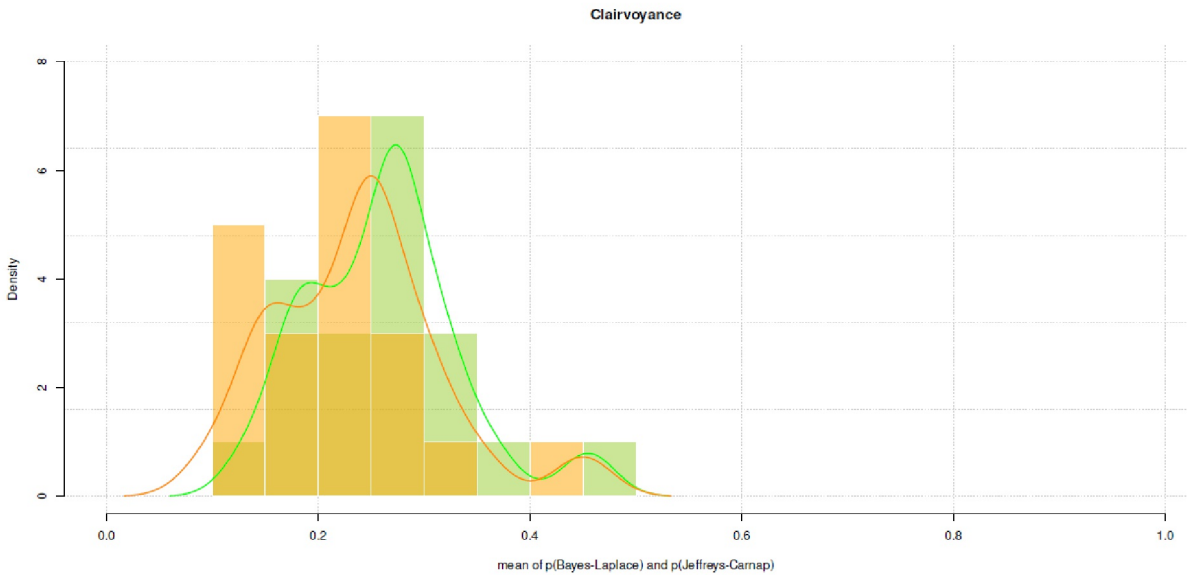


Figura 6.23. *Estudio de la clarividencia (tasas de éxito bayesianos, histogramas)*

Centrémonos en la segunda variable dependiente, la certeza subjetiva (véase también la figura 6.24). Mientras que la hipótesis clara respecto a la clarividencia es que, en el mejor de los casos, hay valores atípicos, pero que no hay ningún efecto en todas las personas de la prueba, no hay ninguna tesis respecto a la seguridad subjetiva. Esto se investiga de forma puramente exploratoria mediante EDA y las conclusiones sólo se refieren a la generación de hipótesis para futuros estudios. Por certeza subjetiva se produjo una evaluación del propio juicio del 0% al 100% por intento de clarividencia. La figura 6.24 muestra que no existe una correlación entre la capacidad de clarividencia y la certeza subjetiva. Por el contrario, existen diferencias entre los grupos de tratamiento y de control en lo que respecta a la certeza subjetiva. El grupo de tratamiento muestra valores significativamente más altos, sin que esto se corresponda con una realidad objetiva. Sin embargo, los datos no sugieren diferencias significativas en la dispersión.

```
# subjective confidence
> summary(hell$subjsicher)
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.0    25.0  40.0   41.6   50.0  100.0
```

Las estadísticas resumidas muestran que se cubre toda la gama, es decir, del 0% al 100% (mediana = 0.4, x = 0.416, s = 0.2118). En general, los sujetos de tratamiento están más convencidos de sus supuestas capacidades de lo que justifica su rendimiento clarividente, y muestran una menor una dispersión menor que el grupo de control.

```
> with(hell, tapply(subjsicher, treat, summary))
$C
Min. 1st Qu. Median Mean  3rd Qu. Max.
0.0  20.0    30.0   37.8  50.0    100.0
$T
Min. 1st Qu. Median Mean  3rd Qu. Max.
0.00 30.00   50.00  45.42 60.00   100.00
> with(hell, tapply(subjsicher, treat, sd))
C        T
21.39389 20.31027
```

Una certeza subjetiva típica del 25% sería adecuada y cercana a la realidad. La mediana y la media están próximas entre sí y, por tanto, son un 15% más altas de lo esperado. Por tanto, la certeza subjetiva inflada en un factor de 40. Se vuelve más interesante si diferenciamos la certeza subjetiva según los grupos de tratamiento

```
> with(hell, tapply(subjsicher, treat, summary))
$C
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.0  20.0    30.0   37.8 50.0    100.0
$T
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.0  30.0    50.0   45.4 60.0    100.0
> with(hell, tapply(subjsicher, treat, sd))
C      T
21.394 20.310
```

y trazamos adicionalmente, una vez diferenciada según la capacidad de clarividencia (véase la Fig. 6.24 izquierda) y una vez a lo largo de la división en los dos grupos (véase la Fig. 6.24 derecha):

```
# boxplots
par(mfrow=c(1,2))
with(hell, boxplot(subjsicher ~ differ.TF, col="bisque",
       notch=TRUE, main="subjective confidence"))
with(hell, boxplot(subjsicher ~ treat, col="bisque",
       notch=TRUE, main="subjective confidence"))
```

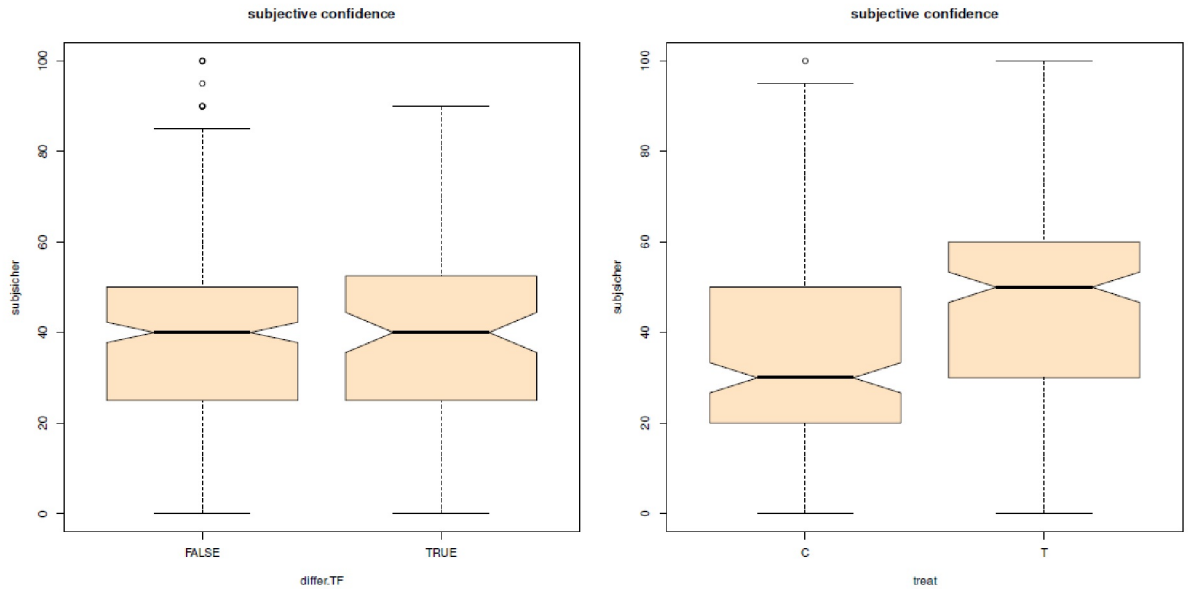


Figura 6.24. *Estudio de la clarividencia (certeza subjetiva, boxplots)*

Numéricamente, se aprecia una diferencia media del 20% (mediana) y del 7.6% (media) entre los niveles del factor grupo de tratamiento y de control del diseño, con valores más altos para el tratamiento. En relación con la certeza experimentada de la probabilidad de acertar, el factor varía del 30% para el grupo de control al 50% para el tratamiento. Así pues, el valor para el tratamiento aumenta en un factor de 2/1.2 = 1:66. Los gráficos también indican una ligera interacción de los factores diseño y resultado de la clarividencia. Los valores de certeza subjetiva para FALSO (= adivinar mal) están más próximos para el grupo de tratamiento y el de control, mientras que los valores para VERDADERO están significativamente más alejados. Esto

significa que para el tratamiento la certeza subjetiva, que de todos modos es mayor que para el grupo de control, aumenta para las respuestas correctas, mientras que para el grupo de control tiende a disminuir de FALSO a VERDADERO. ¿Es esto un indicio de clarividencia implícita, es decir, que los sujetos en la condición de tratamiento aumentaron intuitivamente su certeza cuando acertaron? Y a la inversa, ¿el grupo de control desarrolló algún tipo de habilidad negativa como el miedo a sus propias capacidades, pero aún así mayor que la probabilidad de acertar? Difícilmente – nada de esto es una prueba y en este punto dejamos que los lectores encuentren las respuestas correctas por sí mismos a lo largo de las discusiones en este libro. Como pistas damos algunas palabras clave: probabilidad de acertar, efectos experimentales, efectos perceptivos, muestreo y replicación.

Volvemos al gráfico de interacción, esta vez coloreado por persona y en función del experimento (véase la Fig. 6.25).

```
with(hell, interaction.plot(Upn, trial, subjsicher, col=rainbow(20),
                    bty="n", legend=TRUE,
                    main="Interaction plot\nsubjective confidence",
                    pre.plot=grid()))
```

El paquete R `lattice` abre algunas variantes interesantes de gráficos que permiten trazar por separado por subgrupos, algo así como `xyplot()` o `bwplot()`. Otros gráficos de interacción son útiles para examinar bien los datos. En la Figura 6.25 ya se observa que, del grupo de control, es poco probable que las personas 9 y 20 hayan respondido adecuadamente al caso. La persona 9 representa la tendencia a responder siempre lo mismo y exactamente la probabilidad de acertar del 25%; y la persona 20 probablemente se permitió una pequeña broma al aumentar sucesivamente la certeza subjetiva siempre en un 5%. Ambas se eliminan posteriormente del conjunto de datos, ya que en realidad se espera una respuesta auténtica que surge de la situación. Esto es poco probable para ambas personas, por lo que la eliminación del conjunto de datos parece justificada. Los lectores interesados pueden reproducir los análisis anteriores para un conjunto de datos reducido sin las personas 9 y 20. La eliminación es la siguiente

```
# remove person 9 and 20
hell2 <- hell[-which(hell$Upn %in% c(9,20)),]
```

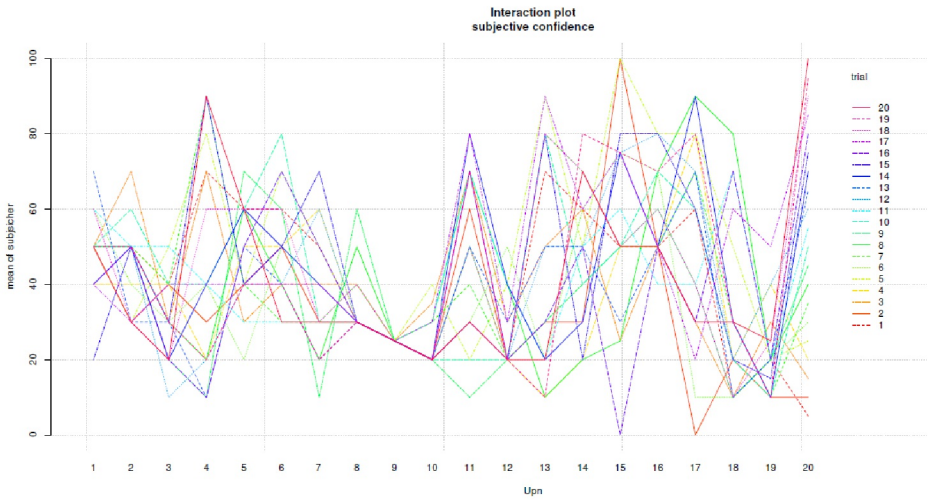


Figura 6.25. *Estudio de la clarividencia (certeza subjetiva, gráfico de interacción)*

Sin embargo, esto no cambia sustancialmente los gráficos. No se da el caso de que los grupos estén completamente separados. Las diferencias descritas entre FALSO y VERDADERO con respecto a la certeza subjetiva se dan en ambos grupos. La suma de los dos grupos da como resultado la situación antes mencionada. Nos quedamos con el conjunto de datos completo "`hell`".

Una prueba bayesiana de los resultados del experimento de clarividencia tiene el siguiente aspecto: en primer lugar, es necesaria una inteligente Prior. Si trabajamos siguiendo a Studer (1996b) con una distribución pre-priori o distribución uniforme, obtenemos

```
# Bayes test success rate
hell.sum <- with(hell, table(differ))
si <- hell.sum["WAHR"]
Ni <- sum(hell.sum)
hell.mean.BL <- round(((si + 1) / (Ni + 2)),dig)
hell.mean.JC <- round(si/Ni,dig)
```

con la salida de la probabilidad posterior media

```
> hell.mean.BL
WAHR
0.239
> hell.mean.JC
WAHR
0.238
```

por lo que en cada caso p = 0.24, lo que acerca aún más la proporción 95/305 = 0.238 a la probabilidad de la tasa. La Figura 6.26 muestra las dos curvas y los estadísticos de resumen, incluidos los IDH de las distribuciones posteriores. Para distinguir mejor la solución de Bayes-Laplace frente a la de Jeffreys-Carnap según Studer (1996b), ampliamos la sección correspondiente en la salida gráfica:

```
# Binomial probs (Studer 1996 paper)
steps <- 1000
theta <- seq(0,1,length.out=steps)
pbl.res <- pbl(theta=theta, si=si, Ni=Ni, loga=TRUE, reexp=TRUE)
pjc.res <- pjc(theta=theta, si=si, Ni=Ni, loga=TRUE, reexp=TRUE)
sN.ME.res <- data.frame(pbl.res, pjc.res)
head(sN.ME.res)
# to see differences better
plot.bl.jc(theta, sN.ME.res=sN.ME.res, si=si, Ni=Ni,
           filling=FALSE, sele=c(0.17,0.35))
```

Se añaden a esto los valores numéricos de ambas soluciones:

```
> sN.ME.post.summary <- sN.post.su(Ni=Ni, si=si)
OUTPUT posterior results and HDI
Bayes-Laplace and Jeffreys-Carnap
$res
ID si Ni BL (mode) JC (mode) BL (mean) JC (mean)
NA 95 400      0.24       0.24      0.24       0.24
BL (sd) JC (sd) BL (var) JC (var)
  0.021   0.021  0.00045  0.00045
$hdi.BL
      rn 0.69% 0.95% 0.99%
lower NA 0.22  0.20  0.19
upper NA 0.26  0.28  0.29
$hdi.JC
      rn 0.69% 0.95% 0.99%
lower NA 0.22  0.20  0.18
upper NA 0.26  0.28  0.29
```

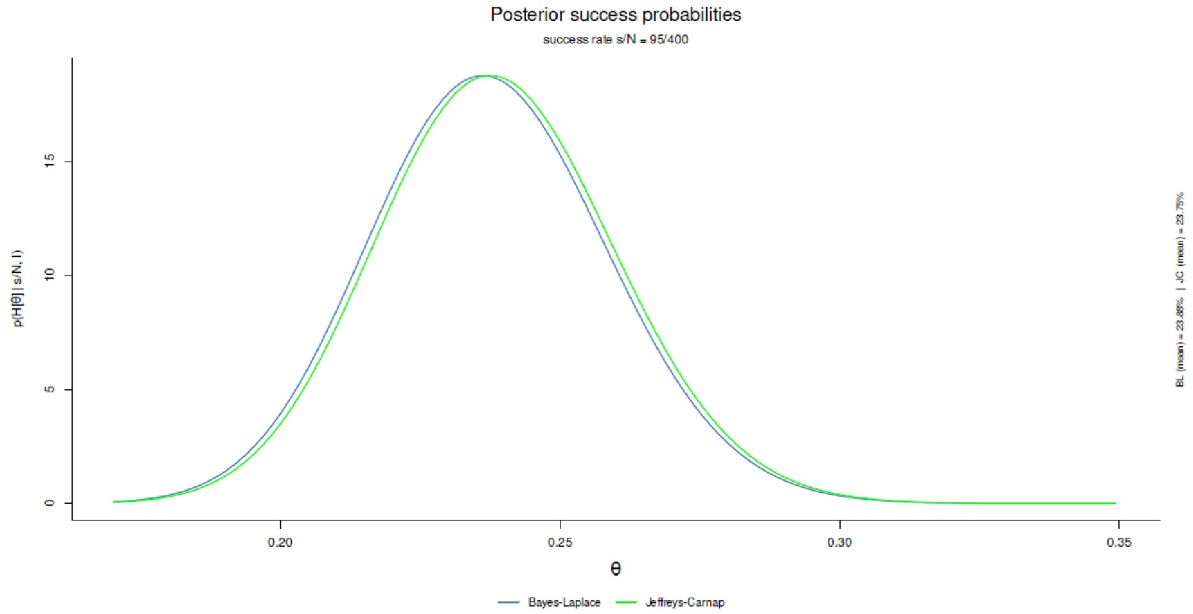


Figura 6.26. *Estudio de la clarividencia (probabilidad de adivinación bayesiana)*

Tanto sobre la simple clarividencia, incluso el amplio intervalo del 99% deja relativamente poco margen para una capacidad sustancial de clarividencia sin descartarla categóricamente, lo que teóricamente tampoco es posible a no ser que trunquemos la distribución a la fuerza. Sin embargo, la probabilidad de que esto ocurra es prácticamente insignificante. Podría ser tan alta como la de un vaso roto que vuelve a recomponerse: teóricamente posible, quizá, pero prácticamente tan improbable que ningún ser humano lo verá jamás.

Una Prior informada es algo diferente, pero cualitativamente muy emocionante. Por lo tanto, la examinaremos en detalle. En primer lugar, consideremos que no hay que probar una hipótesis nula exacta. Probar una hipótesis nula exacta no parece tener mucho sentido en las ciencias sociales (Meehl, 1967). Quizá tenga sentido examinar las constantes físicas con tanta precisión, pero en las ciencias sociales debe asumirse un rango fundamental de variación, precisamente el concepto ROPE (Kruschke, 2015b) o la idea comparable que subyace a las pruebas de equivalencia de la estadística clásica (véase el capítulo 4.4.9). Existe un rango que se considera la *no-existencia* de un efecto y que se puede derivar por motivos de contenido. Si sólo existe una estimación puntual y no se examinara una estimación puntual juntos con una zona de incertidumbre, la masa a priori se situaría completamente en un único valor, en este caso el 25%, lo que parece poco realista, ya que en la práctica no existen efectos nulos al 100%. Hay propuestas para colocar el 50% de la masa a priori en la estimación puntual que se va a investigar y distribuir el otro 50% por igual entre el resto del rango de distribución. Este enfoque – aunque conceptualmente comprensible -- parece igualmente inadecuado para la cuestión que nos ocupa. En primer lugar, no estamos trabajando en condiciones infinitas en las que exactamente el 25% tendría sentido poner el 100% de la masa a priori sobre ella. Pero en el infinito ya no se calcula. Repartir el 50% de la masa por igual al resto requeriría una justificación previa sustantiva de por qué no se toman desviaciones justificadamente. Tal justificación no se nos ocurre. El hecho de trabajar en un contexto no infinito obliga naturalmente a limitar gran parte de la masa prior a un ámbito estrecho según los presupuestos, pero no a un valor singular. Es una cuestión de plausibilidad, y en la clarividencia ésta comienza con la desviación de la probabilidad de adivinación, que en este caso es del 25%. Un ajuste diferente requiere una probabilidad de adivinación diferente y éste es el punto de partida para nosotros.

¿Qué argumentos serían necesarios para fijar el intervalo a priori en un valor distinto del 25%? Aparte de la "creencia" en la clarividencia, que esperamos esté anclada empíricamente, es difícil elegir un valor distinto de la probabilidad de adivinación sin una justificación empírica. ¿Debería ser el 50% o incluso el 67.9% o el 87.22%? No es sólo que haya pocas razones a favor de estos valores, sino que hay igualmente

pocas en contra, incluso si suponemos a priori la existencia de la clarividencia. Una posibilidad sería derivar una Prior de estudios previos. A falta de tales trabajos empíricos previos (de hecho, existe una vasta colección de estudios más o menos buenos sobre la percepción extrasensorial, la mayoría de los cuales se limitan a la lectura de la mente), lo que se necesita es un modelo de un mecanismo de funcionamiento de la percepción extrasensorial y una buena derivación de cómo obtener de él un valor para una Prior. Por cierto, lo mismo se aplica a la neurociencia cuando se trata de la relación entre las áreas cerebrales activas y la experiencia de lo que nos hace humanos. Si no se trata de tejidos y órganos y de actividades en ellos, sino de la experiencia, la sustancia de la neurociencia mengua enormemente, ya que la conclusión causal implícita "localización de la actividad cerebral = calidad de la experiencia" no suele estar justificada, puesto que esto no es nada fácil de investigar y no puede asumirse simplemente y sin cuestionamientos como validez a ojo de buen cubero. La ciencia ejerce una función de modelo; y una investigación de la percepción extrasensorial requiere un buen modelo, del mismo modo que una investigación del comportamiento de aprendizaje de las personas mayores requiere un buen modelo.

Partiendo de un modelo ficticio según el cual la *clarividencia es posible en principio*, ahora habría que justificar ¿por qué debería surgir en las condiciones mencionadas: en un seminario y al adivinar las cartas, por así decirlo, en un contexto cotidiano? En el modelo, la situación experimental por sí sola implica un cierto control de las personas sobre esta presunta habilidad, a saber, mostrarla en el momento *t*. Esto por sí solo es una suposición bastante grande, porque el modelo ya implica que la clarividencia es una capacidad controlable como contar ovejas o girar la cabeza de izquierda a derecha. Otra suposición podría afirmar que tales presentimientos sobre acontecimientos o información *sólo se dan en situaciones extremas* y pondrían en cuestión el modelo de control, por ejemplo, si uno se preguntara en 1912 si podría billete para el Titanic o si sería mejor embarcar en el siguiente barco hacia América. Sin embargo, un modelo así sería difícil de poner a prueba sólo por razones éticas. Además, desde el punto de vista estructural, probablemente sea diferente si se trata de adivinar cartas ocultas como si se tuvieran unos ojos más o si posiblemente se está tomando una decisión existencial de forma inconsciente, lo que, en principio, podría haberse reconstruido a partir de diversas informaciones más una buena parte de "intuición", lo que ciertamente no funciona con la simple adivinación de cartas. En consecuencia, podríamos preguntarnos si ambos objetos de investigación, a saber, la adivinación de cartas por un lado y la cuestión de embarcar en el Titanic por otro, están investigando lo mismo o no. La relación de la intuición y la percepción extrasensorial por sí sola merecería sin duda el lanzamiento de un gran proyecto de investigación, si uno está interesado y dispuesto a investigarla con imparcialidad, y sin ninguna opinión fija preconcebida de que los fenómenos extrasensoriales existen o no. La intuición, como procesamiento inconsciente de la información para tomar decisiones a las que el yo consciente no llega fácilmente, sí tiene solapamientos con áreas que en realidad están reservadas al esoterismo.

Como puede verse, la argumentación se mueve en áreas que ya conceptualmente divergen muy rápidamente y son difíciles de captar. La clarividencia no consiste en leer información de una pizarra. Si se tratara de eso, probablemente muchas personas leerían primero los números de la lotería de la semana siguiente y las empresas de lotería cerrarían en muy poco tiempo. Lo mismo ocurre con los casinos, las casas de apuestas y las previsiones meteorológicas. Y esta última, incluso con métodos matemáticamente complejos, sólo es exacta a muy corto plazo. Nadie puede predecir el tiempo en exactamente 165 días y a la hora. Volviendo a los casinos y los números de la lotería... nada de esto se ha cumplido todavía. Todavía hay muchas organizaciones de juego, por lo que hay que asumir que la mayoría de la gente no puede ser psíquica, o al menos no puede predecir con fiabilidad los resultados de los juegos de azar "a secas". Esto no significa que no haya conexiones entre las personas, que se podrían describir de modo neutral como una *transmisión cualitativa de información* entre las personas y su entorno hasta ahora no investigada, descubierta o comprendida, pero que, sin embargo, parece fantástica. No excluye la posibilidad de que haya individuos en el mundo actual que posean capacidades perceptivas aumentadas. Pero se trataría de excepciones y aquí estamos examinando al público en general normal porque estamos haciendo estadísticas. El análisis de casos individuales no suele requerir estadísticas. En la actualidad, y válido para el público en general, no hay nada que sugiera que las personas simplemente sepan qué cartas se les están mostrando, qué está escrito en un trozo de papel que no pueden ver, qué tiempo hará dentro de quince días y cuáles serán los próximos números de la lotería. No hay que subestimar que la presente prueba experimental sólo funciona con un

conjunto finito de cuatro opciones y que la probabilidad de adivinación por sí sola ya es del 25%. En sentido estricto, la clarividencia pertenece al por qué es posible y no al revés. Si todos supiéramos cosas que en realidad no podemos saber, probablemente seríamos muy estúpidos, porque obviamente no aprenderíamos nada de ello y seguiríamos cometiendo los errores que ya deberíamos saber de antemano gracias a nuestros conocimientos. En este sentido, casi toda la masa de la probabilidad a priori reside en la probabilidad de adivinar y la incertidumbre que la rodea se crea para dar cuenta de las fluctuaciones de la realidad y sus múltiples influencias. No se trata de suavizar la hipótesis de base, sino de definir el margen legítimo que puede conciliarse con la hipótesis nula de que "la clarividencia no existe".. Debido a esta lógica de justificación en realidad se trata de *justificar* la hipótesis nula y *no* de rechazarla. Cualquier otra cosa sería sin sentido.

Esto conduce inevitablemente a la estadística bayesiana. Si no se rechaza la hipótesis nula en términos estadísticos clásicos, no hemos aprendido nada y no podemos cuantificar la incertidumbre asociada en el contexto concreto (pero véase Schimmack, 2015b, en relación a la teoría de Neyman-Pearson).

Pasemos ahora al análisis: se supone, por tanto, que existen fluctuaciones típicas en la probabilidad de tasa del 25% $\pm$ x%. Se supone que esta fluctuación es del 15%, por lo que resulta un 25%, es decir, un intervalo plausible del 10%. Allí se espera el 90% de la masa de la distribución, de modo que el 5% puede estar por debajo y por encima. El 25% se define como la mediana de la distribución a priori y eso corresponde a la probabilidad de adivinar. Debido al hecho de una distribución binomial la distribución beta se presta como Prior conjugado. Aquí el código R (`ptII_quan_Bayes_case_exp-extra-sensual-perception.r`):

```
# determine prior beta
prior.ab <- beta.determine.opt(p=c(0.5,0.95,0.05),
            qua=c(0.25,0.4,0.1), ab.start=c(10,25),
            graph=TRUE, leg1="topright", leg2="right")
str(prior.ab)
hell.prior.res <- do.call("cbind",
            beta.summary(a=prior.ab$res.ab[["a"]], b=prior.ab$res.ab[["b"]]) )
```

La función R `beta.determine.opt()` recibe como entrada los cuantiles 5%, 50% y 95% con los valores de probabilidad *p* de que son: 0:1, 0:25 y 0:4. Esto da como resultado los parámetros a y b de la distribución beta a priori.

```
> # output
> hell.prior.res
        a       b    mode    mean      sd       var
[1,] 4.7752 15.556 0.20595 0.23488 0.091787 0.0084249
> prior.ab$res.ab
    a     b
4.7752 15.5556
> prior.ab$res.ab3
    a     b
5.2974 16.3860
```

La figura 6.27 muestra la distribución prior beta resultante con los valores optimizados para las distribuciones lineal (a = 4.775, b = 15.556) o (alternativamente) cuadrática (a = 5.297, b = 16.386).

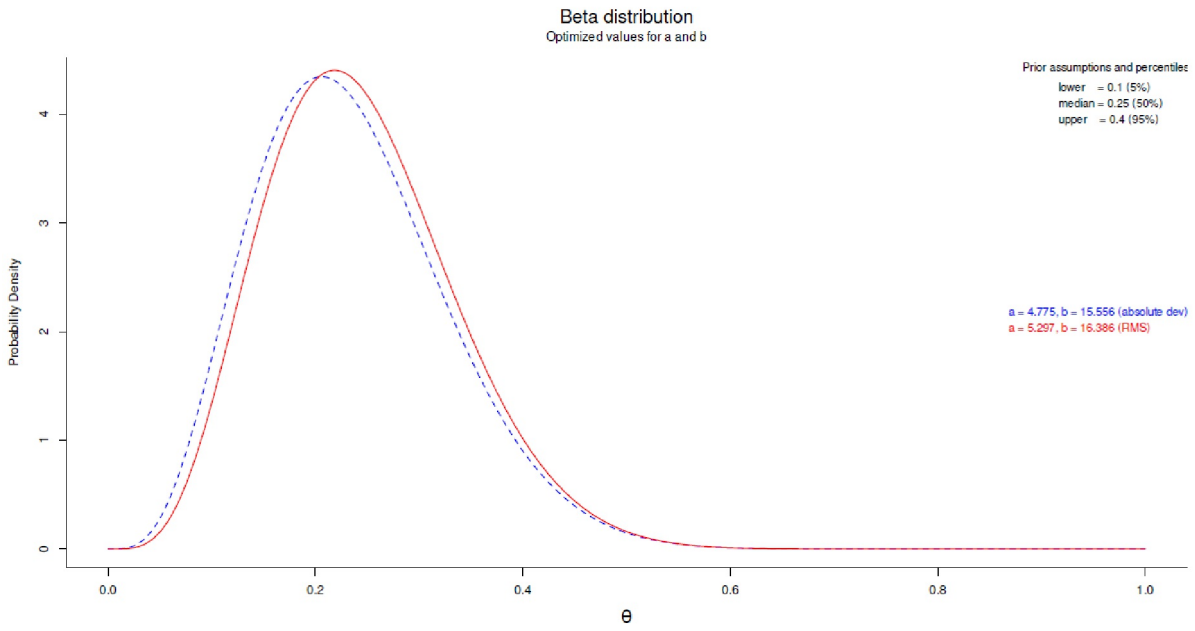


Figura 6.27. *Estudio de la clarividencia (Distribución Beta a priori)*

Con los valores de la desviación lineal, se forma ahora la posterior a partir de la Prior y la Likelihood. Siguiendo a Jaynes (2003, cap. 17.2), se podría considerar si la minimización lineal o la cuadrática es más adecuada desde el punto de vista informático.

```
# likelihood
likeli.ab <- bino.ab.lik(si=si , Ni=Ni)
hell.likeli.res <- do.call("cbind",
          beta.summary(a=likeli.ab[["a"]], b=likeli.ab[["b"]]) )
# create posterior
post.ab <- bino.ab.post(a.prior=prior.ab$res.ab["a"],
          b.prior=prior.ab$res.ab["b"], si=si, Ni=Ni)
hell.post.res <- do.call("cbind",
beta.summary(a=post.ab[["a"]], b=post.ab[["b"]]) )
```

Se puede agregar los datos:

```
v <- c(unlist(hell.prior.res[,c("a","b")]),
       unlist(hell.likeli.res[,c("a","b")]),
       unlist(hell.post.res[,c("a","b","mean","sd")])
)
v <- data.frame(t(v))
colnames(v) <- c("a.prior","b.prior","a.lik","b.lik",
       "a.post","b.post","mean.post","sd.post")
```

Esto da como resultado los estadísticos resumidos de las distribuciones beta para Prior, Likelihood y Posterior

```
> v
    a.prior b.prior a.lik b.lik   a.post b.post   mean.post sd.post
1   4.7752  15.556   96    306    99.775 320.56   0.23737   0.020728
```

y un gráfico final de Prior, Likelihood y Posterior (véase la Fig. 6.27). Ninguno de los valores - y con ello nos referimos sobre todo al intervalo del 99% – no entra dentro de un rango sospechoso de clarividencia. Casi podemos retirarnos: *caso cerrado*.

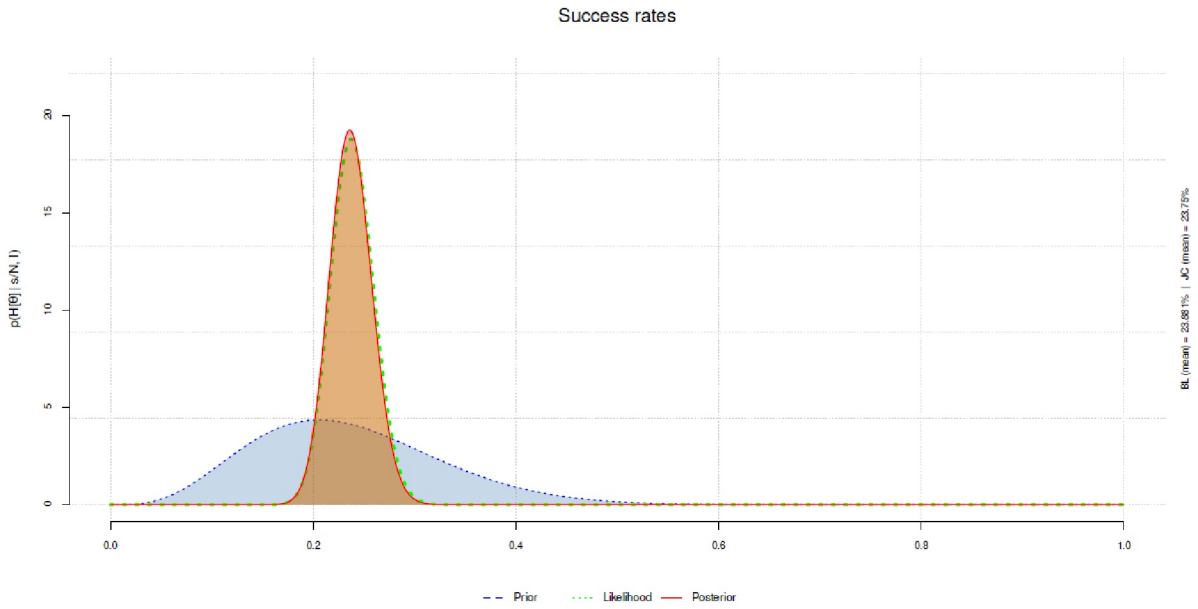


Figura 6.28. *Estudio de la clarividencia (Prior, Likelihood, Posterior, distribución Beta)*

```
beta.triplot(si=si, Ni=Ni, v=v, multiplot=FALSE, musdplot=F,
    plots=c(TRUE,TRUE,TRUE), filling=TRUE)
```

El cálculo de los intervalos de confianza bayesianos (IDH, s. cap. 6.8.4.1) para el 69 %, 87 % y 99 % no muestra nada que no fuera de esperar:

```
      prob     Prior    Likelihood    Post
lower 0.69     0.124011 0.21639      0.21551
upper 0.69     0.308305 0.25951      0.25759
lower 0.87     0.092399 0.20637      0.20572
upper 0.87     0.363991 0.27063      0.26843
lower 0.99     0.043921 0.18584      0.18565
upper 0.99     0.488011 0.29491      0.29212
```

La tabla de contingencias no muestra ninguna sorpresa:

```
> # base contingency table
> hell.tab <- with(hell, table(treat,differ))
differ
treat FALSO TRUE
C       150   50
T       155   45
```

Pasemos ahora a las diversas pruebas bayesianas que parecen interesantes a efectos de demostración sobre el conjunto de datos. El porcentaje de aciertos podría examinarse con `bayes.binom.test()` del paquete R `BayesianFirstAid`. Contrastaremos esto con la prueba clásica de χ^2 más adelante.

```
> bino1 <- bayes.binom.test(95,95+305,cred.mass=0.9, comp.theta=0.25)
> summary(bino1)

Data
number of successes = 95, number of trials = 400
Model parameters and generated quantities
theta: the relative frequency of success
x_pred: predicted number of successes in a replication

Measures
        mean     sd     HDIlo   HDIup  %<comp %>comp
theta   0.239    0.021   0.205    0.274  0.701  0.299
x_pred 95.655  11.967  74.000  113.000  0.000  1.000
'HDIlo' and 'HDIup' are the limits of a 90% HDI credible interval.
'%<comp' and '%>comp' are the probabilities
of the respective parameter being smaller or larger than 0.25.

Quantiles
        q2.5%   q25%    median      q75%     q97.5%
theta   0.199   0.224    0.238        0.253   0.281
x_pred 73.000 87.000   95.000     104.000  120.000
> hdi(bino1$mcmc_samples, credMass=0.8)
        theta x_pred
lower 0.21215  79
upper 0.26606 109
attr(,"credMass")
[1] 0.8
```

Así pues, cabe preguntarse si existen diferencias de clarividencia entre los grupos de diseño. Aparte de que no hay ninguna hipótesis razonable al respecto, los resultados obtenidos hasta ahora muestran que un grupo sería particularmente bueno y el otro particularmente malo de modo que el diseño equilibrado acabaría de nuevo en el rango de los índices. Los IDH no sugieren esto. Tampoco lo hacen Posterior Odds, que nos dicen algo sobre las probabilidades de ser mayores o menores que un umbral crítico (véase también el capítulo 6.8.4.2 sobre ROPE).

```
> # posterior odds
> mean(unlist(bino1$mcmc_samples[,"theta"]) > 0.21)
[1] 0.91987
> mean(unlist(bino1$mcmc_samples[,"theta"]) > 0.21 &
unlist(bino1$mcmc_samples[,"theta"]) < 0.27)
[1] 0.84733
```

Podemos comprobar eso utilizando `contingencyTableBF()` del paquete `BayesFactor` de R. Se selecciona como modelo `SampleType="indepMulti"` y las filas se definen con `fixedMargin="filas"` como constantes con respecto a la suma de las filas (= tamaño de los grupos), pero no las columnas (= número de respuestas correctas e incorrectas). Comparamos el resultado con la prueba clásica de χ^2:

```
> # classical binomial chisquare test
> chisq.test(hell.tab)
Pearson's Chi-squared test with Yates' continuity correction
data: hell.tab
X-squared = 0.221, df = 1, p-value = 0.64
> bf0 <- contingencyTableBF(hell.tab, sampleType="indepMulti",
+          fixedMargin="rows")
> 1/bf0
Bayes factor analysis
--------------
[1] Indep. (a=1) : 7.937155 ±0%
Against denominator:
```

```
Alternative, non-independence, a = 1
---
Bayes factor type: BFcontingencyTable, independent multinomial
```

El resultado deja pocas dudas sobre la independencia del diseño y la clarividencia. El tratamiento no influyó ni positiva ni negativamente en los resultados de clarividencia de este grupo. Con el fin de investigar el asunto por otros medios – sin esperar nuevos hallazgos – pasamos a otras pruebas. La clarividencia relacionada con el diseño puede probarse como prueba de dos muestras con `BESTmcmc()` del paquete R `BEST` o mediante `anovaBF()` del paquete `BayesFactor` de R. En primer lugar, la vista clásica con el conocida `t.test()`,

```
> with(hell.res, t.test(WAHR[treat=="C"], WAHR[treat=="T"],
+' var.equal=FALSE))
Welch Two Sample t-test
data: WAHR[treat == "C"] and WAHR[treat == "T"]
t = 0.681, df = 17, p-value = 0.5
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1.0485    2.0485
sample estimates:
mean of x mean of y
    5.0       4.5
```

y, a continuación, una variante sencilla según Gill (2007) en la aplicación de un artículo de Gönen, Johnson, Lu y Westfall (2005) para el factor de Bayes de la hipótesis nula de que las dos distribuciones no difieren. La prueba se basa en el estadístico de prueba *T* de la diferencia de medias estandarizada, en el que se basan en última instancia todos los factores de Bayes a este respecto. El factor de Bayes se calcula a partir del cociente de dos modelos. Por un lado, se trata de la densidad para el estadístico de prueba *T* con grados de libertad *df* y parámetro de no centralidad *ncp = 0* frente a la densidad para el estadístico de prueba *T* con grados de libertad *df* así como localización *ncp* y parámetro de escala *pv*. Los dos últimos parámetros se calculan a partir de los supuestos previos sobre la diferencia de medias estandarizada, es decir, la media δ y la varianza σ^2 (Gönen, Johnson, Lu & Westfall, 2005, p.253). La función de R `t.test.BF()` contiene las implementaciones correspondientes. Según los autores, el procedimiento pertenece al procedimiento de Bayes, más bien subjetivo (véase el capítulo 6.5.2), ya que espera que los investigadores conozcan su contexto y utilicen la información previa correspondiente para realizar una estimación plausible de la Prior. Wang y Liu (2016) critican este enfoque y, entre otras cosas acusan al enfoque del peligro de la paradoja de Lindley-Bartlett (véase el capítulo 4.4.14.2) y proponen una aplicación matemática puramente objetiva. A su vez, Gönen, Johnson, Lu y Westfall (2019) proporcionan contraargumentos invalidantes, no solo examinando los factores de Bayes más comunes para la comparación de grupos de ramas en un estudio de simulación para la clasificación correcta, sino también proporcionando argumentos por los que una elección de la Prior sensible al contexto es preferible a una elección situacionalmente no plausible pero objetiva. Un ejemplo que citan es, por ejemplo, la elección de la Prior del factor de Bayes en el estudio de Bem (2011a). A continuación se presentan los datos del experimento:

```
> # simple Bayesian t-Test with prior believes clairvoyance
> means.TC <- with(hell.res, tapply(WAHR, treat, mean))
> contr <- hell.res$WAHR[hell.res$treat=="C"]
> treat <- hell.res$WAHR[hell.res$treat=="T"]
> do.call("rbind", lapply(list(contr=contr,treat=treat),
+          function(x) c(N=length(x), summary(x), SD=sd(x),
+          VAR=var(x), fivenum2(x))))
        N Min. 1st Qu. Median Mean 3rd Qu. Max.  SD   VAR
contr  10 3    4.00     5     5.0  5.00    9   1.8257 3.3333
treat  10 2    3.25     5     4.5  5.75    6   1.4337 2.0556
        minimum lower-hinge median upper-hinge maximum
contr   3        4             5       5          9
treat   2        3             5       6          6
```

```
> cohensd(s1=contr, s2=treat)
d|mean     sd d|pooled    sd d corrected|N<50
-0.30460   -0.30460       -0.27676
> hell.bf.t <- t.test.BF(mean.delta=0.5, sigma.delta=2,
+              samp1=contr, samp2=treat)
Bayesian two-sample t-Test
Gönen, Johnson, Lu and Westfall (2005)
BF01 = 3.62
BF10 = 0.277
Test BF01 > BF10 = TRUE
Test BF01 < BF10 = FALSE
Result = BF in favor of H0
NOTE: BF01 favors zero difference | BF10 favors a difference
```

El factor de Bayes sugiere que los grupos no difieren entre sí. Y ahora la variante bayesiana completa, empezando por `BESTmcmc()`. El script R contiene otros análisis que se basan en `BESTmcmc()`.

```
> # BEST by Kruschke
> hell.best <- with(hell.res,
+ BESTmcmc(WAHR[treat=="C"], WAHR[treat=="T"]))
> summary(hell.best)
           mean   median mode  HDI% HDIlo HDIup compVal %>compVal
mu1        4.920  4.909  4.930 95  3.557  6.27
mu2        4.534  4.540  4.567 95  3.423  5.62
muDiff     0.386  0.379  0.330 95 -1.350  2.16   0          67.5
sigma1     2.004  1.895  1.708 95  0.935  3.29
sigma2     1.634  1.541  1.398 95  0.862  2.65
sigmaDiff  0.370  0.333  0.263 95 -1.214  2.02   0          69.5
nu        33.101 24.426  8.245 95  1.082 92.09
log10nu    1.355  1.388  1.460 95  0.548  2.08
effSz      0.210  0.211  0.209 95 -0.693  1.15   0          67.5
> plot(hell.best) # not shown
> # group differences
> meandiffhell.best <- hell.best$mu1 - hell.best$mu2
> sigmadiffhell.best <- hell.best$sigma1 - hell.best$sigma2
> mean(meandiffhell.best)
[1] 0.3898676
> mean(sigmadiffhell.best)
[1] 0.3649971
```

A continuación se realiza el análisis con el paquete `BayesFactor` de R. Aquí se plantea la situación de que podríamos examinar inmediatamente un modelo más complejo que, además del tratamiento, incluya la certeza subjetiva y la medición repetida por persona o la persona en sí. Dado que esta investigación es exploratoria, es legítimo probar específicamente sin más hipótesis profundas, así como intentar comprender el conjunto de datos. Si encontráramos algo, primero necesitaríamos una buena hipótesis y explicación, y una investigación de replicación o seguimiento basada en esto. El número de experimento y la persona pueden tratarse como *random effects* (Pinheiro & Bates, 2009; Gelman & Hill, 2007). Dado que en la estadística bayesiana los parámetros se consideran básicamente como variables aleatorias, la estadística bayesiana no tiene la separación de efectos fijos y aleatorios que tiene la estadística clásica. Gelman, Carlin, Stern y Rubin (2003, p.383 nota 1) señalan a este respecto

> „The terms 'fixed' and 'random' come from the non-Bayesian statistical tradition and are somewhat confusing in a Bayesian context where all unknown parameters are treated as 'random' or, equivalently, as having fixed but unknown values."

Los resultados de `generalTestBF()` se basan en factores de Bayes y, a diferencia de `anovaBF()` y `lmBF()`, permite la combinación de predictores continuos y categóricos. Sin embargo, no se trata de una

solución posterior bayesiana completa. Nos limitamos a probar todos los modelos. Como estamos trabajando de forma exploratoria y esto es más para demostrar la implementación, eso está permitido. La variable `Upn` (persona) sirve como *random effect*. Además, el tratamiento `treat`, la certeza subjetiva `subjsichersum` como el valor medio de cada persona a través de sus ensayos, y la interacción de tratamiento y certeza subjetiva se utilizan como predictores (*fixed effects*) en el modelo.

```
# BayesFactor R-Code
# Upn as factor
hell.res$Upn <- factor(hell.res$Upn)
# create a prob value of TRUE vs. (TRUE+FALSE)
hell.res$probres <- with(hell.res, WAHR/(FALSCH+WAHR))
# create a mean subj confidence value
hell.res$subjsichersum <- with(hell, tapply(subjsicher, Upn, mean))
hell.res
str(hell.res)
# all models - does make sense only
# for exploratory tasks, not in general!
hell.res.BF <- generalTestBF(probres ~ treat * subjsichersum,
        data=hell.res, whichRandom="Upn",
        noSample=FALSE,whichModels='all')
hell.res.BF
```

con la salida

```
Bayes factor analysis
--------------
[1] treat :                                          0.46807 ±0%
[2] subjsichersum :                                  0.41546 ±0%
[3] treat:subjsichersum :                            0.51613 ±0%
[4] treat + subjsichersum :                          0.19148 ±0.86%
[5] treat + treat:subjsichersum :                    0.23266 ±0.88%
[6] subjsichersum + treat:subjsichersum :            0.27329 ±0%
[7] treat + subjsichersum + treat:subjsichersum :    0.1286 ±1.09%
Against denominator:
Intercept only
---
Bayes factor type: BFlinearModel, JZS
```

Los potentes factores de Bayes nunca prueban contra un modelo de Intercept only ; y nada contradice nuestras expectativas. También podríamos trazar el modelo y las Posteriores:

```
plot(hell.res.BF)
# reduce error
hell.res.BF.re <- recompute(hell.res.BF, iterations=5e+5)
hell.res.BF
hell.res.BF.re
# posterior
nsamps <- 1e+5
hell.res.samps <- posterior(hell.res.BF, index=4, iterations=nsamps)
head(hell.res.samps)
summary(hell.res.samps)
plot(hell.res.samps, col="violetred3")
```

El resultado global no indica que los sujetos de este estudio fueran clarividentes y que el tratamiento pudiera haber influido en ello de algún modo. Si el ruido blanco distrajo a este grupo, al menos no fueron externamente peores que la probabilidad de adivinar, sino que se sintieron más seguros de haber acertado en sus respectivos ensayos. Esto podría compararse con la medición post hoc de la *concentración subjetiva*

durante el experimento y sería una tarea para los lectores. Del mismo modo, está la actitud personal subjetiva hacia la percepción extrasensorial *antes y después* del experimento. Hay material suficiente para otras pequeñas hipótesis.

```
> # short and rough overview for further work
> # attitude towards ESP pre + post
> with(postsubj, table(ASWpre,ASWpost))
ASWpost
ASWpre -1  0 1
-1      5  0 0
 0      0 11 1
 1      0  0 3
> # attitude towards ESP pre + post across groups
> with(postsubj, ftable(treat, ASWpre,ASWpost))
            ASWpost -1 0 1
treat ASWpre
C      -1            3 0 0
        0            0 5 0
        1            0 0 2
T      -1            2 0 0
        0            0 6 1
        1            0 0 1
> # perceived concentration across groups
> with(postsubj, ftable(treat, Konzentration))
Konzentration 10 20 30 50 70 80 90 100
treat
C              4  0  1  0  0  1  2  2
T              0  2  1  2  2  0  3  0
> with(postsubj, plot(treat, Konzentration))
```

A pesar del claro resultado, repetimos nuestro mantra incluso aquí: *necesitamos una buena teoría, replicación y variación*. Sin embargo, podría ser que el tiempo de arranque con k = 20 ensayos fuera demasiado corto, lo que entonces – véase más arriba – requeriría su propia lógica de razonamiento, así que tenemos que repetirnos: *Necesitamos un modelo que pueda predecir todo esto de antemano (¡!). Y necesita efectos convincentes*.

Tarea 6.6: Clarividencia y certeza subjetiva

La figura 6.24 muestra una clara diferencia entre los grupos a lo largo del diseño para la variable t *seguridad subjetiva*. ¿Con qué se relacionan? ¿Diferencias en las medias? ¿Diferencias en las desviaciones estándar? ¿A ambas? ¿Ninguna de las dos, sino algo más? La tarea aquí para los lectores dedicados sería a lo largo de los R-análisis anteriores repetirlos con la certeza subjetiva como variable dependiente. Los predictores pueden combinarse a voluntad – estamos trabajando aquí de forma exploratoria – siempre que nadie piense después que han demostrado algo. Si se pueden encontrar uno o más efectos, la siguiente tarea sería formular una buena teoría desde la psicología al respecto. Esto debería ir seguido de una breve descripción de un estudio de seguimiento que tome los efectos encontrados y explicados post-hoc como punto de partida para derivar hipótesis formalmente correctas. El script R

`ptII_quan_Bayes_case_exp-extra-sensual-perception.r`

contiene más sugerencias de implementación.

6.8.1.6 Crítica al diseño: el estudio de Bem

Como digresión final, añadimos nuestra propia crítica teórica y metodológica al estudio de Bem, ya que los estudios críticos y los intentos de réplica citados en la literatura son sólo estadísticos y se centran en la recogida de datos y apenas reflejan el trabajo teórico subyacente o el diseño y la selección de la muestra. Le damos la vuelta, dejamos de lado las estadísticas en su mayor parte, ya que esto ha sido suficientemente discutido por otros autores, y nos centramos en el trabajo teórico así como en la selección de la muestra.

Para empezar: se puede acusar a Bem de enormes deficiencias en su trabajo teórico, que bien podría haber subsanado con el esfuerzo adecuado y un poco de inversión en su diseño. Describiríamos su estudio, que se limita casi exclusivamente a una explicación evolutiva, como bastante libre de teoría, por lo que los posibles efectos apenas tendrían poder explicativo. Su selección de la muestra también es desfavorable y probablemente se deba a que en las universidades es fácil reclutar sujetos de estudio que no tengan que ser capaces de hacer nada especial. Esto no es cierto para muchos experimentos en psicología (social).

Trabajaremos principalmente de forma cualitativa en los argumentos y nos concentraremos en el trabajo teórico que subyace al estudio de Bem. Idealmente, esto sería una teoría que explica y predice diversos fenómenos parapsicológicos, describe su desarrollo y proporciona técnicas para el desarrollo, así como las intervenciones para ayudar a las personas con, por ejemplo, problemas en este contexto. Para un estudio empírico, la teoría debería dar margen suficiente para derivar hipótesis precisas y falsables que puedan formularse de forma coherente. Adoptamos explícitamente el punto de vista de que tales fenómenos son posibles y reuniremos pruebas de que el estudio de Bem todavía atrae críticas masivas bajo esta postura no neutral. La parapsicología en sí es un campo antiguo y el término fue acuñado por el filósofo y psicólogo Max Dessoir (1867-1947) en 1889 y se siguió desarrollando en el siglo XX en la dirección de la investigación experimental. Se suele distinguir entre percepción extransensorial y telequinesia, es decir, aumento de la percepción de los sentidos e intervención mediante trabajo mental. Aunque a menudo se ridiculiza esta disciplina, cuenta con una larga tradición y sigue estudiándose y desarrollándose en forma de asociaciones. En Friburgo de Brisgovia existe un centro de asesoramiento parapsicológico para la población, que fue apoyado económicamente y con éxito por el estado de Baden-Wurtemberg durante casi 30 años. Está dirigido por el físico y psicólogo Walter von Lucadou (2012). Los fenómenos con los que se trabaja allí no son relevantes aquí, en primer lugar porque se producen de forma espontánea e incontrolada y, en segundo lugar, porque parecen ser una externalización de los problemas y conflictos humanos, que se trasladan al entorno físico externo de forma casi comparable a la psicosomática. Esta parece ser una explicación razonable según von Lucadou (ibid.) para explicar, incluso sin una explicación científicamente sólida, por ejemplo inexplicables ruidos, movimientos de objetos, etc. en un contexto con sentido. Con la solución de estos problemas realmente humanos, estos fenómenos inexplicables suelen desaparecer. El hecho de que haya otros fenómenos que no queden cubiertos por esta explicación no es relevante para el debate posterior del estudio de Bem. En resumen, sin embargo, parece existir una necesidad tanto de información como de asesoramiento en este campo. En nuestra opinión, sólo esto es suficiente para investigar el tema de forma seria y científica sin descartarlo preconcebidamente como un disparate o creer ciegamente en él o incluso idealizarlo. En sentido estricto podríamos preguntarnos por qué se investiga la teoría física de *strings*, ya que ni siquiera es accesible experimentalmente... la libertad de la ciencia consiste en investigar lo que a cada investigador le parece digno. Hasta aquí la situación inicial.

Como marco de referencia teórico para evaluar el estudio de Bem, elegimos el sistema de enseñanza práctica del Buda, Siddhata Gotama, que al fin y al cabo tiene más de 2500 años y ha tenido y tiene millones de practicantes desde entonces. Este sistema tiene la ventaja de cubrir, casi de pasada, una amplia gama de fenómenos que comúnmente caen en el ámbito de la parapsicología o el esoterismo. Elegimos deliberadamente hacer esto para mostrar que el estudio de Bem tiene enormes deficiencias, incluso si se adopta un punto de vista que permita todos los fenómenos posibles, y esto incluye bastantes. El sistema doctrinal budista entiende que los fenómenos supuestamente sobrenaturales son reales, pero no se considera que tengan una función o relevancia significativa para la vida. Al contrario, se consideran un obstáculo en el camino hacia la iluminación, ya que pueden distraer e inflar el ego. En contraste con la actitud de que tales fenómenos no existen, se adopta así la actitud de que estos fenómenos existen, pero que no son deseables. Y si se producen, deben afrontarse con gran cautela.

La enseñanza budista no trata de la práctica de estos fenómenos o habilidades, algo que la mayoría o incluso todos los enfoques esotéricos no pueden pretender. En ellos, se trata principalmente de la consecución de habilidades especiales (por ejemplo, prever el futuro, leer los pensamientos, influir en la materia a través de la mente) o de experimentar experiencias sensoriales especiales (por ejemplo, viajes astrales, estados extáticos). En su forma original, tal y como se recoge en el Tipit.aka (U Ko Lay, 1995), el sistema de enseñanzas de Buda no es de naturaleza religiosa o sectaria, sino que se entiende como una ciencia práctica de la mente y la materia, exclusivamente para mostrar a la gente el camino hacia la iluminación, la experiencia de un estado aparte de la mente y la materia. Sólo se enseñan técnicas para explorar el propio fenómeno mente-cuerpo, que pertenecen al campo de la concentración, especialmente *A*n*a*p*a*na-Sati (= observación de la respiración que entra y sale de forma natural para concentrar la mente) y el desarrollo de la sabiduría, Vipassan*a* (= meditación de la sabiduría a través de la observación de las propias sensaciones físicamente perceptibles) (V.R.I., 1993; Goenka, 1997). Esto se complementa con recomendaciones claras para una vida ética (S*i*la, cinco reglas para laicos) y con meditación para desarrollar la bondad amorosa (Mett*a*) con el fin de compartir los logros propios con todos los demás seres. Las técnicas de *A*n*a*p*a*na-Sati y Vipassan*a* constituyen el núcleo y la base de prácticamente todas las técnicas practicadas (por ejemplo, Kabat-Zinn, 2005) y evaluadas (por ejemplo, Anderssen-Reuster, 2007; Studer, 1998) que se conocen hoy en día bajo el término meditación de atención plena. Básicamente, se trata "sólo" de respirar sin hacer nada y observar las sensaciones corporales tangibles sin reaccionar a ellas con apego (rechazo, codicia/querer tener) y desarrollar una actitud benevolente hacia todos los seres. Esto suena sencillo, y en teoría lo es, pero en la práctica es increíblemente difícil "no hacer nada", como puedes comprobar fácilmente por ti mismo: Intenta concentrarte durante un minuto (o más) en tu respiración natural, sin dejar que tu mente divague.

Si lo intentas, serás un poco más realista sobre tus propias capacidades. Y eso ni siquiera nos lleva al terreno de la meditación de la sabiduría, sólo de la concentración. Bem extrae su legitimidad de metaestudios (Honorton & Ferrari, 1989), que informan de un efecto pequeño pero consistente con una muestra total de N = 50.000 personas y k = 309 experimentos realizados por m = 62 investigadores diferentes. Merecería la pena examinar este metaestudio por separado, pero no proporciona un marco teóricamente sólido para todos los estudios reportados, ya que Bem sólo se refiere a él empíricamente, pero no teóricamente.

A continuación examinamos las explicaciones teóricas antes de que Bem pase a las empíricas. Bem se une a los estudios (Radin, 1997, 2006) que trabajan con la percepción subliminal (índices fisiológicos, excitación emocional ante estímulos negativos o eróticos, etc.) y sus efectos de excitación eran medibles antes de la presentación de estos estímulos. Normalmente, los resultados de los experimentos según Bem (2011a, p.408, véase también la Tabla 1) tenían la siguiente estructura:

> „As expected, strong emotional arousal occurred when these images appeared on the screen, but the remarkable finding is that the increased arousal was observed to occur a few seconds before the picture appeared, before the computer had even selected the picture to be displayed."

Sería interesante investigar cómo se relaciona exactamente la percepción inconsciente con la capacidad de clarividencia. En realidad, la clarividencia debería ser un acto consciente y reflexivo – es decir, una acción – si se quiere considerar una habilidad. Trasladarla al ámbito del inconsciente parece problemático, porque aquí falta el eslabón para transformar el acto pasivo de la percepción inconsciente en una capacidad y una acción conscientes, que las personas puedan dominar. Si anticipamos a continuación el marco del sistema de enseñanza de Buda, este sistema modela las capacidades extraordinarias como capacidades humanas raras pero conscientes y no como inconscientes y por lo tanto incontrolables. Esto no significa que tales fenómenos, dependiendo del contexto, del nivel y grado de realización alcanzado, de la modalidad de los sentidos, etc., no puedan manifestarse también de forma espontánea y, por tanto, incontrolable. Pero a la larga, la orientación debe y tiene que ser el dominio del fenómeno, es decir, el control consciente, que corresponde a la definición de una acción (Heckhausen, 1989). Falta, pues, una integración teórica de la conciencia y la inconsciencia, así como de la percepción, la toma de decisiones y la acción en. Bem señala además (2011a, p.408)

> „The major theoretical challenge for psi researchers is to provide an explanatory theory for the alleged phenomena that is compatible with physical and biological principles. Although the current absence of an explanatory theory for psi is a legitimate rationale for imposing the 'extraordinary' requirement on the evidence, it is not, I would argue, sufficient reason for rejecting all proffered evidence a priori."

Aun estando de acuerdo con él en que la falta de una teoría coherente no debe obstaculizar el empirismo, es falso que no existan enfoques teóricos sobre el tema que no puedan ser fructíferos para la ciencia. No hay ninguna razón por la que esto sólo deba provenir de la física o la biología. Parece como si Bem nunca hubiera tratado el tema *teóricamente*. Desde luego que los hay, aunque mantendríamos las distancias con la mayoría de los intentos de explicación, ya que nos suenan demasiado *esotéricos* y a menudo tienen teorías demasiado poco específicas. Pero eso no significa que no se puedan probar de forma falsable. Si un planteamiento teórico, por *absurdo* que sea, es en principio falsable, se puede investigarse empíricamente. Que eso siempre tenga sentido es otra cuestión. Eso no lo juzgamos aquí. En cambio, sostenemos que la ciencia debe tener libertad para hacer cosas absurdas. Bem, sin embargo, no aborda su evitación de las explicaciones esotéricas. Pero eso tendría mucho sentido – aunque sólo sea porque el esoterismo es un gran negocio y está rodeado de muchos mitos – de milagros y mitos como curaciones espontáneas, poderes omnipotentes, etc. Desde el punto de vista social, pues, habría sido sumamente útil analizar las explicaciones esotéricas y examinarlas para ver si son empíricamente accesibles en absoluto. En cambio, Bem se centra en la sección de teoría en unos pocos (meta)estudios empíricos (véase más arriba) y no deriva sus hipótesis teóricamente de ninguna manera. Practica así un empirismo casi sin teoría, que sólo consideramos legítimo si no hay intentos teóricos de explicar los fenómenos investigados. No es éste el caso. Probablemente haya incluso demasiados intentos de explicación, por ejemplo desde el campo del esoterismo. Pero ni siquiera éstos son necesarios para una teoría útil. En cuanto a la teoría, Bem no está haciendo su trabajo - una omisión fatal.

Por ejemplo, Bem podría haber citado al físico Burkhard Heim (1925-2001), quien además de las cuatro dimensiones del espacio-tiempo afirmó la existencia de dos dimensiones cualitativas adicionales, una estructural y otra teleológica, que supuestamente tienen el potencial de ser utilizadas para explicar y predecir, entre otras cosas, fenómenos parapsicológicos, ovnis, viajes en el tiempo, etc. (von Ludwiger, 2013). Así que habría algunas opciones, que no sonarían del todo acientíficas sólidas para formular un edificio hipotético comprobable. Bem describe su "teoría" (ibíd., p.6) de la siguiente manera:

> „The presentiment studies provide evidence that our physiology can anticipate unpredictable erotic or negative stimuli before they occur. Such anticipation would be evolutionarily advantageous for reproduction and survival if the organism could act instrumentally to approach erotic stimuli and avoid negative stimuli."

Más adelante profundizaremos en la explicación evolutiva no falsable. Sin embargo, todo esto es más bien poco, ya que la formulación y justificación de la hipótesis no permite afirmar nada en absoluto sobre la cantidad y la dirección exacta de los efectos según la situación y no predice nada aparte de la suposición de un efecto evolutivo general. Todos los conocimientos de Bem sobre el tema sólo salen a la luz en la discusión, pero ésta es *post hoc* y, por tanto, de un valor explicativo extremadamente limitado. ¿Por qué no utilizó este *a priori* para dar una visión teórica bien fundamentada que incluyera hipótesis derivadas? Si Bem hubiera mirado un poco más de cerca en las tradiciones asiáticas, como el sistema de enseñanza budista, habría sido fácil encontrar que el término ESP = ExtraSensorial Perception es erróneo y engañoso, ya que el término "extra" implica que se está añadiendo un nivel adicional de percepción, como simplemente cambiar el ADN y añadirle una hebra o dar a nuestros sentidos otro sentido para hacer posibles cosas que comúnmente no lo son. Pero eso es innecesario.

Según el *sistema budista de enseñanza*, todas las experiencias, no sólo las de la parapsicología, tienen lugar enteramente *dentro* de nuestro propio fenómeno mente-cuerpo. Así, la mente se considera un sentido separado junto a los otros cinco más el cuerpo material. Así pues, se supone que la mente y el cuerpo no son idénticos en principio, pero son muy interactivos entre sí, pero nunca algo independiente, fuera o por encima de nuestros seis sentidos. Todos los fenómenos, tanto materiales como inmateriales, están sujetos a la misma

ley natural, el Dhamma. El uso de un término lingüístico como ESP es, desde este punto de vista, simplemente erróneo e implica cosas que nunca son redimibles – a saber, la experiencia de algo *fuera* de nuestros sentidos por medio de nuestros sentidos. Esta definición no es falsable y por lo tanto no es investigable científicamente y por lo tanto es un caso para el esoterismo. Si la definición ya es errónea, no habrá más operacionalización a nivel empírico. En inglés, existen términos mucho mejores como "clairvoyance" o "2nd sight", que generalmente pueden traducirse como "clarividencia". El primero tiene raíces francesas y significa "ver con claridad" y el segundo sugiere que existe al menos otra perspectiva además de la percepción cotidiana normal que se puede adquirir. Ambos términos nos parecen mucho más apropiados para enmarcar el fenómeno. Bem, por su parte, escribe constantemente sobre la percepción extrasensorial.

Esos efectos a los que comúnmente se hace referencia como clarividencia, etc. son, desde un punto de vista budista, o bien consecuencias (*cámmicas*) de acciones previas (= principio causa-efecto, Pa-Auk Tawya Sayadaw, 2009/2012) que ahora están siendo "cosechadas" como "fruto" (por ejemplo, acumuladas a lo largo de muchas existencias del fenómeno mente-cuerpo) y manifestándose, o bien son un efecto secundario de estados concentrativos muy profundos de meditación y de progreso en la meditación, las jh*a*nas, es decir, estados concentrativos de absorción (Pa-Auk Tawya Sayadaw, 1999/2019; Solé-Leris, 1994). No se trata en absoluto de fenómenos perceptivos extrasensoriales, ya que todo lo potencialmente experiencial tiene lugar siempre dentro del propio fenómeno mente-cuerpo a lo largo de los seis sentidos, sin excepción. Se trata más bien de una sensibilidad muy fuertemente aumentada en la percepción sensorial y en las capacidades mentales, lo que cualitativamente crea una connotación completamente diferente. A la inversa, con respecto a las posibilidades humanas de influencia, existe una sensibilidad comparable para influir en niveles muy finos de la realidad, de modo que los efectos también aparecen en el nivel manifiesto. Estos últimos serían entonces los fenómenos que parecen efectos sobrenaturales, pero que en realidad no lo son. Todo tiene lugar en nuestro mundo, no fuera de él.

No se trata explícitamente de percibir o provocar algo fuera de lo natural, sino que a través de una sensibilidad muy grande se pueden percibir o incluso cambiar cosas que siempre están ahí, pero que generalmente no somos capaces de percibir ni, por tanto, ni asignarles atención ni significado, lo que probablemente sea mejor así. Imagínese poder ver con clarividencia o mirar en la mente de otras personas todo el tiempo. Esto requeriría un alto nivel de procesamiento e integración y una estabilidad ética del que la mayoría o casi todas las personas no serían capaces. Por eso, esto daría lugar a una sobrecarga masiva en la vida cotidiana, con los correspondientes excesos negativos a nivel de actuar (abusos y similares). Lo mismo cabe decir de las capacidades que en un sentido más amplio modifican la materia. Así que si estas habilidades existen, son de origen natural y la naturaleza lo ha dispuesto bien para que generalmente no tengamos nada que ver con ellas. Aunque estos fenómenos no tienen un fin en sí mismos, el camino para desarrollarlos es sin embargo extremadamente largo y agotador. ¿Qué dice esto sobre la distribución de tales cualidades en la población general? Probablemente que prácticamente no se encuentra a nadie que realmente posea tales capacidades y que las domine ampliamente, no sólo de forma rudimentaria o espontánea, sino de forma controlada y repetible. Pero es precisamente esto último lo científicamente interesante. Estimamos que las fuerzas de efecto esperadas en la población general son prácticamente nulas.

La perspectiva presentada tiene varias consecuencias, de modo que Bem – si le hubiera hecho caso – podría haberse ahorrado por completo su estudio de esta forma, ya que representa, en el mejor de los casos, ilusiones pero no ciencia con base teórica y no respalda ninguna conclusión razonable que incorpore y combine cualitativamente la información contextual existente. Como experimento mental, seguimos tomando en serio el sistema de enseñanza de Buda como modelo de la realidad y vemos qué otras predicciones pueden derivarse de él.

- Si una mayor capacidad de percibir hasta este punto es consecuencia de grandes inversiones en el pasado o un subproducto de elevados estados meditativo-concentrativos en el presente, entonces se deduce para el presente: La probabilidad de habilidades parapsicológicas sustanciales en personas normales de una muestra como la reunida por Bem es prácticamente nula. Ignoramos, por razones pragmáticas, los detalles complejos y de interrelaciones colectivas complejas y para nosotros inmanejables, porque no verificables, y los efectos sobre el desarrollo de capacidades en el individuo. Probablemente sólo muy pocas personas en el mundo encajan

en estas categorías. Por tanto, una Prior sería ponderado con masa extrema a cero y una dirección de prueba unilateral.

- En el sistema de enseñanza budista todos los fenómenos parapsicológicos en principio tienen su lugar y parecen fantásticos. Pero también se enseña que la exhibición pública de tales artes es insalubre porque infla el ego y por lo tanto obstruye masivamente el camino hacia la iluminación – por lo que debe evitarse bajo cualquier circunstancia. Por lo tanto, suponemos que las personas que, debido a altos estados meditativo-concentrativos, pueden poseer selectivamente habilidades de esta naturaleza son significativa-mente menos propensas a participar seriamente en tal estudio o a estar en un estudio de este tipo o que muestren algo en absoluto. La práctica de la meditación prolongada e intensiva y el estilo de vida asociado a ella sugieren que estas personas, por lo general, no pretenden demostrar que tienen tales capacidades. Esto sería ciertamente si estas personas estuvieran iluminadas desde el punto de vista budista y por lo tanto no podrían tener ya consecuencias negativas para ellos mismos. Bem no sería capaz de reconocer a estas personas, porque no existe ningún instrumento de selección para ellas. Personas que desarrollan tales habilidades como espontaneamente, no podrian pertenecer a esta categoría. Es posible que muestren tales habilidades, pero no se espera que sean muy comunes. Pues la idea que subyace es que cuando tales habilidades (incluso espontáneamente) se muestran, en algún momento del pasado debe haber tenido lugar una gran inversión en su desarrollo. Estas habilidades no existen de la nada y sin historia.
- Ambas categorías de causa – cámica y meditativa – no permiten elegir específicamente el tipo de perceptividad aumentada, es decir, si uno puede adivinar las cartas, ver en la mente de otras personas, predecir el futuro, mentir, materializar cosas, etc., como si se tratara de un supermercado donde se pueden elegir tales artes a voluntad. Aunque al Buda histórico Siddhata Gotama se le atribuyó prácticamente todas estas facultades, al mismo tiempo se menciona en el Tipitaka, que no eran un fin en sí mismas en el sistema del Buda, sino que eran, en el mejor de los casos, subproductos. En el Tipitaka, sólo se pueden encontrar muy pocos acontecimientos individuales en la vida del Buda en los que supuestamente demostró habilidades especiales. Pero como se señala muchas veces que el Buda dejó claro que estas habilidades no eran deseables porque se trata de la iluminación y no de hazañas, es cuestionable si estos pasajes del Tipit.aka son auténticos, es decir, si el Buda demostró alguna vez realmente algo extraordinario ante un auditorio. El Tipitaka describe qué habilidades tenían sus monjes, monjas y seguidores laicos (Nyanaponika & Hecker, 1997). Por ejemplo, *A*nanda, el primo y secretario de Buda tenía la capacidad de la memoria fotográfica incluso antes de su iluminación y tras la muerte de Buda pudo establecer la base del primer concilio budista, ya que era capaz de reproducir exactamente todas las enseñanzas de Buda que había escuchado. Esto dio origen al Tipit.aka. Las razones para desplegar estos poderes se describen en la cita anterior, y esto es bastante típico de los relatos budistas: No se muestran, y si se muestran, es de forma inconsecuente por los iluminados y para ayudar de alguna manera a los demás en su camino. Así, el Buda tenía la capacidad de ver en la mente de los demás, y la utilizaba esto para dar instrucciones sobre la práctica de la meditación según el caso. Ésta fue la segunda de las últimas instrucciones que el Buda dio a *A*nanda inmediatamente antes de su muerte. Estos relatos no hacen sino demostrar que opiniones serias sobre las capacidades parapsicológicas que desempeñan un papel en ciertas tradiciones humanas y que son tradiciones humanas y que no son arbitrarias ni caprichosas. Por otra parte, desde el punto de vista budista, tales capacidades no se distribuyen simplemente por igual. Es decir, no se va y se elige qué habilidades pueden manifestarse, si es que alguna persona llega incluso a desarrollarlas. Muchísimos discípulos del Buda lograron profundas percepciones sobre la naturaleza de su propio fenómeno mente-cuerpo y alcanzaron el estatus de arahat – pero no poseían ningún poder parapsicológico significativo. En resumen eran personas sabias, pero no podían doblegar leones con la mente, ni elevarse por los aires ni hacer otras cosas extraordinarias como ver los números de la lotería y otras cosas sin importancia. Sin embargo, Bem no trata de estas habilidades espirituales en el sentido estricto de la palabra, es decir, no se trata de sabiduría. Todos estos puntos no aumentan de ninguna manera la Prior, la probabilidad de que incluso una sola persona de la muestra de Bem sea capaz de mostrar estas habilidades de forma controlada y reproducible.
- Sacar los números de la lotería, adivinar naipes o predecir estímulos eróticos o negativos no se encuentran entre las habilidades mencionadas como relevantes en las tradiciones budistas o en el Tipitaka, por decirlo amablemente. El Tipitaka como legado del Buda, organizado en discursos doctrinales, preceptos para monjes y monjas, y el trabajo escolástico sobre la mente y la materia comprende decenas de miles de páginas de instrucciones muy serias sobre la conducta de la vida. Por tanto, es muy informativo sobre lo que se considera relevante e irrelevante en el sistema budista. De este modo, Bem no sólo ha hecho extremadamente desfavorable la selección de muestras y el trabajo teórico. Ha elegido como objeto de estudio algo que no es muy relevante, sino que probablemente requeriría un gran esfuerzo para desarrollarlo. La práctica común de adivinar los naipes o los resultados de los juegos de azar, o las operacionalizaciones utilizadas por Bem de "estímulos eróticos y estímulos negativos" siguen formando parte, sin duda, de los comportamientos indignos de la Orden en la actualidad. En el caso de la demostración de habilidades por parte de monjes o monjas, la

> expulsión (inmediata) de la orden sería presumiblemente la consecuencia. Se podría imaginar que la predicción de acontecimientos futuros peligrosos parece más aceptable, pero esto se contradice con declaraciones del Buda de que se trata de aceptar el aquí y ahora para utilizarlo como base para configurar el futuro de forma sana y no para huir de él.

Podríamos detenernos ya en este punto. Bem no hizo bien su investigación – ninguna hipótesis, operacionalizó el objeto de estudio de forma muy desfavorable y selecciona mal su muestra, recurriendo simplemente a los estudiantes de su universidad esperando que sus capacidades psíquicas sean suficientemente dimensionadas. Sin embargo, los argumentos anteriores no sólo se oponen a que las personas sean ampliamente clarividentes, sino que demuestran que la teoría, la metodología y la selección de la muestra utilizadas eran inadecuadas. No es posible deducir de ello si el fenómeno existe o no. Muy al contrario, ni siquiera hemos llegado tan lejos. Según el enfoque bayesiano podemos deducir a priori que la masa de la distribución se encuentra en cero o, en el experimento, en la probabilidad de conjetura. Permitimos una cierta vaguedad mínima en torno al exterior, que subyace a la propia conjetura. Si, en estas circunstancias, sale un resultado empíricamente "clásicamente significativo", entonces debemos ser escépticos y cuestionar la metodología y el análisis de los datos.

El intento de Bem de vender teóricamente la clarividencia como un *rasgo evolutivo* también apunta a un pobre trabajo teórico. Porque o bien la gente ya tiene estas habilidades, en cuyo caso habrían estado usándolas durante mucho tiempo (=validez ocular), porque lo que es técnicamente posible, la gente lo pone en práctica primero, sin importarle las consecuencias (por ejemplo, la energía nuclear, la tecnología de vigilancia, la deforestación de la selva tropical), etc.) o no lo tienen. Si el experimento hubiera resultado realmente positivo y las pruebas empíricas hubieran sido un éxito fuera de toda duda, incluyendo réplicas, seguiría siendo un mal trabajo teórico, ya que las explicaciones evolucionistas simplemente funcionan post-hoc y no permiten ninguna predicción real, ni siquiera intervenciones. La utilidad queda así severamente limitada, a menos que se trabaje como antropólogo geólogo, etc., que trabajan con periodos de tiempo más largos y no pueden trabajar de otra manera. Para ellos, por supuesto, las explicaciones evolutivas no son un mal trabajo teórico. Estos largos periodos de tiempo, por otra parte, no desempeñan un papel significativo en las ciencias de la vida.

Una suposición de Bem, que los sujetos de prueba de repente tienden inconscientemente hacia un nivel de clarividencia mas alto – sobre la base de unas pocas fotos desnudas – parece ingenua. Hoy en día estamos constantemente rodeados de algo más que fotos de desnudos (publicidad, internet, ...) y en la muestra de Bem (EE.UU., universidad) las fotos de desnudos ya no representan nada especial. Debería en vista de la publicidad, etc. y las personas ligeras de ropa que los acompañan (si sigue siendo así), o por la fácil disponibilidad de dicho material. ¿O ciertos grupos de personas (consumidores de porno) no han desarrollado capacidades clarividentes extremadamente altas? ¿Lo han hecho? No lo parece. Las teorías evolucionistas siempre adolecen del hecho de que ni son comprobables ni falsables, pero como se mencionó anteriormente, proporcionan explicaciones post-hoc, que suenan bastante razonables, pero generalmente no permiten deducciones directas para la actualidad. Los plazos de la evolución son sencillamente demasiado largos y las consecuencias de la evolución no se desarrollan en el contexto del experimento. Por lo tanto, los resultados son sólo irrelevantes explicaciones post-hoc sin mucha sustancia. Dado que las sociedades de lotería siguen existiendo, es probable que la clarividencia simplemente no se haya desarrollo evolutivo, de lo contrario estas sociedades y todos los casinos de juego casinos habrían tenido que cerrar hace mucho tiempo, porque más dinero es sin duda una ventaja reproductiva. Este punto de vista tiene cierta validez aparente, pero no representa una evidencia empírica.

Examinemos la cuestión con cifras ficticias para hacernos una idea de estas dimensiones. En 2011, según Wikipedia (2019d), había unos 7.000 millones de personas en todo el mundo y de las cuales 312 millones vivían en Estados Unidos, el país en el que se realizó el estudio de Bem en la Universidad de Cornell en el estado de Nueva York (Bem, 2011a). Suponiendo lo anterior aceptando los tipos anteriores, entonces – valores supuestos – 90% de los que han adquirido capacidades excepcionales a través de la *meditación* no están dispuestos a participar en un estudio, y el 80% de los que han desarrollado sus capacidades e*spontáneamente* están explícitamente dispuestos. Además está el hecho de que el nivel de capacidad no

será del 100% en todas partes, es decir, las habilidades no siempre pueden reproducirse con exactitud. Simplifiquemos y fijemos la tasa de replicación en un 60% bastante alto, es decir, en el 60% de los casos una persona puede replicar sus habilidades de forma fiable, o el caso es que el 60% de las personas pueden replicar sus habilidades. Este último caso es más fácil de tratar a nivel de personas que de sucesos. Además, partimos del supuesto poco realista de una distribución equitativa, es decir, que esas personas estén distribuidas por igual en el mundo. Siendo realistas, en el caso de la meditación, esas personas deberían encontrarse en centros particulares, por ejemplo, monasterios, centros de meditación, ashrams, etc. El otro grupo del caso manifestación espontánea podría distribuirse teóricamente por igual. No está claro si la edad desempeña un papel. En la muestra de Bem la edad es muy limitada, ya que se trata de estudiantes universitarios (undergraduates). Además, se supone que ambos grupos forman cada uno el 50% de la población total de personas con capacidades. Estimemos el número de personas con habilidades parapsicológicas a un nivel profesional relativamente alto k = 1000 personas en todo el mundo, de las que tomamos una cifra simplificada del 60%, es decir, personas que pueden reproducir sus habilidades de forma fiable. Consideramos todo esto como un límite superior. Además, consideramos que la distribución por sexos es equitativa y no le damos más vueltas. El resultado es la siguiente imagen para todo el mundo en 2011, si *k* representa el número de personas en todo el mundo con las con las habilidades estudiadas, *r* por la capacidad de replicar, *BMed* por la voluntad, la habilidad en el caso de la *meditación*, *Bspon* para la disposición a mostrar *espontáneamente* la habilidad y *q* para la proporción de *casos de meditación frente a casos espontáneos* en relación con la población, formulado como probabilidad. Los supuestos anteriores conducen a *k = 1000*, *r = 0.6*, *BMed = 1 - .9*, *Bspon = 0,8* y *q = 0.5*

(`ptII_quan_Bayes_Bem-estudio-aspectos.r`):

Caso Meditación nMed = (k · r) · q · BMed = 30 personas o pMed = 30/7e9 = 4.286e – 09 para reclutar a una tal persona para el estudio;

Caso espontaneidad nspon = (k · r) · q · Bspon = 240 personas o pspon = 240/7e9 = 3.429e – 08 para reclutar a una tal persona para el estudio;

Total ntotal = nMed + nspon = 30 + 240 = 270 personas o ptotal = 270/7e9 = 3.857e–08 para reclutar a una sola persona de los casos CasoMed o Casospon para un estudio.

Globalmente, la probabilidad de reclutar a alguien con esas aptitudes es de *ptotal* = 3.857e- 08, es decir, todas las 1/3.857e-08 = 25 925 925 personas o en todo el mundo *nworld* = 270 personas encontramos una persona - relacionada con 7.000 millones de personas, que encaje con el perfil y esté dispuesta a participar en un estudio de este tipo. Para los EE.UU. esto significa que la probabilidad es de 312e6/7e9 = 0.045, es decir, el 4.46% de la cuota mundial, hace p_{USA} = 0.045 *270 / 312e6 = 3.857e- 08 o (aún) todo 1/ .857143e-08 = 25 925 925 personas en los EE.UU.están dispuesto y tienen las aptitudes para el estudio. La probabilidad de ganar la lotería es 1:13e6 o exactamente `choose(49,6)` = 1 : 13 983 816, por lo que p_{Lotto} = 1/`choose(49,6)` = 7.151124e-08. Así pues, $p_{ratio} = p_{Lotto}/p_{USA}$ = 7.151e- 08/3.857e- 08 = 1.854 es más probable de ganar la lotería que de encontrar a alguien de este grupo objetivo "igualito" en EE.UU. Los lectores interesados pueden modificar ahora el ejemplo numérico con valores más permisivos o conservadores, es decir, cuántas personas se van a contratar en todo el mundo y cuáles son las probabilidades de que estas personas tengan tanto las aptitudes como la voluntad y disposición para participar en un estudio de este tipo. Nuestros valores son ficticios y no pretenden ser válidos. Sin embargo, vemos en ellos cierta plausibilidad para hacernos una idea de las dimensiones del fenómeno, ya que obviamente no todo el mundo a nuestro alrededor las tiene. Los lectores motivados pueden preguntarse cuántas personas han conocido en su vida que hayan tenido las capacidades extraordinarias descritas, y cuántas personas han conocido en su vida. Al menos a nivel subjetivo, esto debería ser convincente. Bem informa de N_{Exp} = 1100 personas sobre k = 9 experimentos realizados en la misma institución (estudiantes universitarios, Universidad de Cornell, Ithaca/ NY, EE.UU.), lo que habla en contra de una extracción aleatoria de la población estadounidense y presumiblemente para mediciones repetidas, es decir, las personas fueron sometidas a pruebas varias veces.

Por desgracia, no hay nada sobre esto en el artículo de Bem (2011a). Si la aleatorización fuera cierta, el azar la probabilidad sería 1 de (1/3.857e -08)/1100 = 23 569.02 personas, es decir, con un tamaño de

muestra de 1100 personas diferentes y una distribución equitativa dada en la población, la probabilidad es 1/23569.02 = 4.243e-05 = 1/choose(23569,1) = 0.004% de encontrar una sola persona. Para redondear todos los individuos de la muestra según los criterios anteriores, necesitamos N_{Exp} = 1100, hay un p_{1100} = 1/choose(23569,1100) ≈ 0%, o expresado logarítmicamente p_{1100} = log(1)-lchoose(23569,1100) = –4440.609. Para comparar, ya n = 10 personas conducen a p_{10} = 1/choose(23569,10) = 6.874e-38 o logarítmicamente p_{10} = -lchoose(23569,10) = -85.57055. Sea como sea, la probabilidad prior de que Bem obtenga una muestra parapsicológicamente adecuda es teórica y prácticamente en un nivel casi cero. Esto no cambiaría mucho aunque partiéramos de 1.000, ahora supongamos 10.000 personas con capacidades excepcionales en todo el mundo. Los lectores motivados pueden calcular los cambios que se producirían. Y ni siquiera hemos tenido en cuenta todas las limitaciones. Por ejemplo, faltan las afirmaciones sobre la modalidad sensorial preferida (véase más adelante), el entrenamiento, el ejercicio, si el experimento es capaz de activar en absoluto las capacidades del individuo, etc.: faltan todos factores que deberían influir significativamente en un estudio de este tipo.

En su trabajo, Bem parte de la base de que todas las personas tienen en principio estas capacidades, pero que se encuentran por debajo del umbral de la percepción, sin embargo particularmente los experimentos sobre percepción subliminal activan estas capacidades, pero probablemente no son directamente conscientemente accesibles, lo que Bem no investiga sistemáticamente. Esta perspectiva lleva a muchas preguntas sin respuesta sobre condiciones, influencias, diferencias interindividuales en cuanto a modalidades sensoriales (véase más adelante), nivel de difíciles de clasificar debido a la falta de fundamento teórico. No obstante, si nos lo tomamos en serio, la argumentación podría continuarse de tal manera que la transfirieran de la percepción inconsciente a la acción consciente. Si se combinan las explicaciones anteriores del sistema doctrinal con las exigencias prácticas de la ciencia y sus criterios, se podría deducir las siguientes características para el fenómeno a investigar "capacidades extraordinarias en los seres humanos", que una buena teoría del fenómeno debe ser capaz de explicar. Podría ser que existen varias condiciones axiomáticas, es decir, condiciones que no son verificables en principio, pero que *deben* aceptarse *a priori* como dadas, ya que pueden constituir la base de una futura validez de constructo. En primer lugar, esto no es un defecto, porque la ciencia se basa fundamentalmente en ciertos axiomas generalmente aceptados que carecen de justificación última – en cambio, la ciencia trabaja con un criterio relativo de verdad, las leyes formales de la lógica, etc.

- Cualquier ser humano puede desarrollar el fenómeno y tiene potencial para hacerlo. Pero no está claro cuánto tiempo necesita el desarrollo y en qué grado se producirá. El fenómeno está en la naturaleza del ser humano y en principio es desarrollable, como el intelecto u otras capacidades. Es de suponer que hay personas superdotadas y otras que lo son mucho menos, como ocurre con todas las capacidades humanas.
- Esto lleva a la interacción común herencia-entorno con factores como los genes y la aptitud, el nivel de socialización y estimulación, así como la calidad de la formación o estímulo y el grado de práctica individual.
- Desde el punto de vista budista, el nivel kámico de las relaciones causa-efecto pasadas, que dejamos de lado por falta de factores comprobables y científicamente reconocidos, pero que no negamos por completo, ya que desempeña un papel importante en el sistema doctrinal budista. No hay razón para no seguir un planteamiento tan seriamente formulado, que ha sido una parte esencial de este sistema de práctica durante hace más de 2600 años.
- El esfuerzo para desarrollarlo es evidentemente muy elevado, por lo que muy pocos pueden ya demostrar el fenómeno de forma reproducible. Por lo tanto, no podemos suponer que la gente lo domina en general, sino que siempre debemos suponer lo contrario. No se trata de habilidades que puedan aprenderse en un curso de fin de semana o en un entrenamiento "al margen" de la vida cotidiana.
- Se distingue entre los casos "frutos de la meditación" y "espontaneidad del fenómeno debido a pasado, esfuerzos no investigables" como los descritos anteriormente.
- Es de suponer que intervienen diversos factores contextuales desconocidos y por especificar en qué forma o cualidad y hasta qué punto se puede demostrar algo.
- Existen distintos grados de realización y niveles entre los que difieren las personas. Esto corresponde con los distintos niveles de realización que tienen las personas con respecto a habilidades generales (por ejemplo, la

música, el arte, el deporte, la cognición, la memoria, la destreza, etc.). Cuando se compara a las personas, esto desempeña un papel importante.

- Las capacidades pueden diferir en función de la modalidad sensorial preferida por cada persona. Puesto que no hay nada más que los seis sentidos, todas las capacidades deben pasar por ellos. Esto significa que las personas pueden ver, oír, saborear, etc., cuando se trata de aumentar la percepción y la recepción de información. También ellas pueden preferir uno o varios de los canales sensoriales como fuente de información, pero no todos por igual ni siempre al mismo tiempo. Lo mismo puede ocurrir con capacidades de intervención, es decir, cuando las personas actúan sobre el mundo de forma extraordinaria. No cabe esperar una distribución igual en el tiempo y las modalidades sensoriales. Eso sería muy atípico en los seres humanos. En consecuencia, sin embargo, es necesario tomar modelar estos diferentes accesos de forma equivalente para integrar la misma calidad pero diferente acceso qua modalidad sensorial limpiamente. Esto parece muy exigente porque el tamaño de las muestras es muy pequeño.
- En resumen, se da la situación de que, en general, se cabe esperar ningún efecto o efectos tan mínimos, que son difíciles o imposibles de detectar. En el caso de aquellos individuos extremadamente raros que (podrían) tener realmente capacidades, los tamaños de los efectos tendrían que ser muy grandes y constantes. Sin embargo, en vista del reducido número, un estudio de este tipo pasa de una metodología estadística a un análisis caso por caso.

No pretendemos que nuestras explicaciones puedan ya explicar mucho. Deben solamente esbozar como experimento de pensamiento que el fenómeno puede explicarse ciertamente con supuestos (axiomáticos), un trasfondo teórico y afirmaciones que pueden ser razonablemente de modo que puedan crearse constructos comprobables empíricamente. A partir de estos esfuerzos debería demostrarse que un modelo general que no tenga en cuenta los factores anteriores yerra el tiro. Sin embargo, es precisamente ese modelo general el que Bem persigue con su muestra, los estudiantes universitarios, en la universidad. El tema en sí es muy interesante; al fin y al cabo, la historia de la humanidad está llena de relatos sobre personas que han tenido y siguen teniendo capacidades extraordinarias. En lugar de avanzar en un tema tan interesante, Bem relega el tema al reino del empirismo sin teoría con datos que apenas parecen serios y una teoría evolutiva minimalista que no se puede probar directamente. Esto tiene como consecuencia que muchos investigadores que podrían estar abiertos para tal tema ahora prefieren elegir un tema menos delicado para perseguir sus intereses por la investigación y para no poner en peligro su reputación. Con ello, Bem no hace ningún bien al tema; al contrario, lo destruye con tal planteamiento.

Hemos adoptado aquí deliberadamente un punto de vista que en principio permite *todas* las posibles capacidades parapsicológicas y no excluye ninguna. E incluso desde este punto de vista, el experimento de Bem parece mal preparado, mal justificado y planificado y con los consiguientes datos no válidos. Podríamos haber formulado la situación de forma mucho más restrictiva, por ejemplo contra el telón de fondo de otros estudios fallidos en el pasado que investigaron similares topicos – que sin duda se pueden encontrar. No se trata de eso. Tomemos el punto de vista anterior – todas las capacidades parapsicológicas son en principio posibles y existen, Bem habría hecho mejor en llevar a cabo un pequeño estudio de casos con personas seleccionadas que han demostrado repetidamente capacidades parapsicológicas y que están dispuestas a demostrarlas. Entonces habría podido observar las capacidades de distinta calidad, modalidad y alcance y examinar lo que realmente aportan los individuos para registrarlas científicamente, documentarlas y buscar explicaciones razonables para ellos. Si este enfoque incluye la anticipación de estímulos eróticos y negativos a lo largo de la percepción subliminal es cuestionable a muy improbable. Una metodología de este tipo estaría orientada a los casos, sería cualitativa, no anclada en la estadística y presumiblemente muy cercana a los experimentos con palomas de Skinner cuidadosamente realizados con muchas repeticiones. Pero también se enfrentaría al problema general de la *replicación* o la *reconstrucción* de los factores contextuales favorecedores u obstaculizadores, o la prueba de la independencia del contexto de tales fenómenos y la cuestión de las diferencias interindividuales. Si Bem llega a un resultado positivo de la prueba sobre la base de estas expectativas previas de cerca de cero, deberíamos dudar de él, no sólo estadísticamente, no sólo de la credibilidad de los datos, sino de todo el diseño y el fundamento teórico. Del punto de vista bayesiano – con una distribución a priori tan significativa, Bem necesitaría resultados empíricos gigantescamente positivos para cambiar la distribución posterior en la dirección de "éxito – hay clarividencia". Cualquier otra cosa sólo conduce a que la influencia de la Prior permanezca como la expectativa la establece. Dablander (2015) resume fielmente la situación general en una entrada de blog:

„We can agree on how much the data support precognition (as quantified by the Bayes factor). However, this does not mean we have to buy it. Extraordinary claims require extraordinary evidence."

Tarea 6.7: Falsificación o no

La tarea para los lectores sería ahora considerar si el experimento descrito ha falsificado completamente el tema de la clarividencia.

6.8.1.7 Factores de Bayes, ¿y ahora?

¿Son los factores de Bayes buenos o malos, como sugieren polémicamente algunos artículos? ¿Se deberían validar siempre los modelos de la estadística de Bayes con factores de Bayes – comparables a una prueba del cociente de Likelihood en la estadística frecuentista – o nunca? Si bien la primera pregunta es difícil de responder – los métodos son tan buenos como se utilicen y se tienen en cuenta sus limitaciones específicas – la segunda pregunta puede responderse con un rotundo "no". Los factores de Bayes no corresponden a un análisis plenamente bayesiano y no pueden sustituirlo cuando se trata de trabajar de forma plenamente bayesiana. Esto no significa que no sean útiles si sus afirmaciones se interpretan correctamente. Para reiterar:

Los factores de Bayes permiten estimar el cambio en las hipótesis a priori sobre los datos a partir de los propios datos, pero sin implicar directamente a los posteriors, aunque las probabilidades de Likelihood se pueden determinar de forma puramente computacional mediante las probabilidades a posteriori y las probabilidades a priori (véanse las Ecs. 6.39 a 6.46 de la p.541). Los factores de Bayes no son probabilidades posteriores, sino probabilidades de Likelihood. Sin embargo, para responder a preguntas de investigación interesan las probabilidades posteriores, y cuando se trata de comparar modelos o similares, las probabilidades posteriores tienen un interés adicional.

Varios ejemplos demuestran este problema de los factores de Bayes. Por ejemplo, Kruschke (2015a) muestra que el factor de Bayes para una prueba de hipótesis nula de la media conduce a un resultado diferente que para el tamaño del efecto. Las razones residen en cómo influye la Prior en los factores de Bayes en cada caso. Por otra parte, la variable posterior de la media y el tamaño del efecto se muestra en el ejemplo de Kruschke como bastante invariante a los cambios en la Prior. Esta no es la única razón por la que Kruschke (2015b, cap. 12; 2013a, 2011a) advierte repetidamente contra el uso rutinario y sin reflexión de los factores de Bayes, especialmente porque contribuyen poco a la estimación de parámetros. Aunque los partidarios de los factores de Bayes podrían argumentar, con razón, que no se deben usar Priores inapropiados – lo que requeriría una definición de lo apropiado y lo adecuado que difícilmente se puede cumplir en todos los casos – Kruschke (2011a) llama la atención sobre el hecho de que los factores bayesianos se pueden usar rutinariamente sin reflexión, especialmente porque contribuyen poco a la estimación de parámetros. Kruschke (2015a) llega a la siguiente conclusión:

„First, the BF [*bayesian factor; adición de los autores*] for the mean (mu) need not lead to the same conclusion as the BF for the effect size unless the prior is set up just right. Second, the posterior distribution on mu and effect size is barely affected at all by big changes in the vagueness of the prior, unlike the BF."

Otro ejemplo común (por ejemplo, Kruschke, 2015d, con código R y JAGS) examina el nivel general de la Posterior y, en consecuencia, la sobreinterpretabilidad de los factores de Bayes. Si la Posterior es baja, es decir, los sucesos tienen solo una probabilidad modesta, ni siquiera los cambios masivos en las expectativas – operados como enormes factores de Bayes – pueden cambiar esta situación, como muestra la crítica al diseño del estudio de Bem. La probabilidad a posteriori sigue siendo pequeña, por ejemplo, porque

determinados sucesos tienen una probabilidad muy pequeña debido a las bajas tasas de base, y esto tiene poco que ver con el hecho de que las expectativas cambien masivamente como resultado de nuevos datos.

Como estudio de caso, consideremos de nuevo la aplicación discreta de la estadística de Bayes para el diagnóstico de enfermedades, que se utiliza en exceso en la literatura (véase el capítulo 6.3.3 para un estudio de caso), y recordemos que el mundo es en realidad continuo y no discreto por naturaleza. Para ello, los ejemplos discretos pueden resolverse simplemente con el teorema de Bayes, una calculadora y sin MCMC.

Así pues, nos interesa saber si alguien tiene una enfermedad dado el resultado positivo de una prueba. La prueba puede ser correcta (= sensibilidad) o falsa (= falsa alarma o falso positivo). La probabilidad a priori de tener la enfermedad (= tasa base) es baja con $p_{enfermedad} = 0.01$. Por tanto, $pno_{enfermedad} = 1 - p = 0.99$ de la población no tiene la enfermedad. La prueba da un resultado binario, es decir, enfermedad sí frente a no o $p_{test.positivo} = 1$ y $p_{test.negativo} = 0$.

La práctica demuestra que la sensibilidad de la prueba es $p_{Sensibilidad} = 0.97$ y la tasa de falsas alarmas $p_{falso.alarma} = 0.05$. La especificidad de la prueba es, por tanto, $p_{Especificidad} = 1 - p_{falso.alarma} = 0.95$ (`ptII_quan_Bayes_BayesFactors_further-remarks.r`).

```
# BF mean and ES different
# M = Model
# D = Data
# prior prob for having a disease
# random draw from population with prior probs
p_disease <- 0.01 #M=1
p_nodisease <- 1-p_disease #M=2
# diagnostic test
# binary outcomes
test.positive <- 1 #D=1
test.negative <- 0 #D=0
# correct detection = sensitivity
p_test.positive_disease <- 0.97 #p(D=1|M=1)
p_test.negative_disease <- 1-p_test.positive_disease #p(D=0|M=1)
# false rate alarms
p_test.positive_nodisease <- 0.05 #p(D=1|M=2)
# specifity
p_test.negative_nodisease <- 1-p_test.positive_nodisease
#p(D=0|M=2)
```

La probabilidad Prior antes de cribar los datos es el cociente de las probabilidades de las tasas base en la población, es decir, la probabilidad Prior de tener la enfermedad frente a no tenerla, independientemente del resultado de una prueba.

```
prior.detvsfa <- p_disease/p_nodisease
```

El factor de Bayes es el cociente entre la probabilidad de tener la enfermedad si la prueba es positiva frente a la de no tenerla si el resultado de la prueba es positivo. Obviamente ya no se trata sólo de las tasas de base, sino de los resultados de las pruebas.

```
BF_detvsfa <- p_test.positive_disease/p_test.positive_nodisease
```

En relación con esto, la Posterior es el producto de las probabilidades a priori por el factor de Bayes (= probabilidades de Likelihood).

```
# posteriors
# p(M=1|D=1) = p(D=1|M=1)p(M=1)/p(D=1)
# p(M=2|D=1) = p(D=1|M=2)p(M=2)/p(D=1)
# posterior odds
```

```
# p(M=1|D=1)/p(M=2|D=1) = p(D=1|M=1)/p(D=1|M=2) * p(M=1)/p(M=2)
# detection vs false alarm
post_detvsfa <- p_test.positive_disease /
                p_test.positive_nodisease *
                p_disease / p_nodisease
```

Así se obtienen los valores de Prior, factor de Bayes y Posterior:

```
> # outputs
> # prior
> prior.detvsfa
[1] 0.01010101
> # BF as incdication of change of expectations
> BF_detvsfa
[1] 19.4
> # posterior
> post_detvsfa
[1] 0.1959596
```

Obviamente, las probabilidades a priori, la Likelihood y las probabilidades a posteriori difieren considerablemente. Aunque de los datos se desprende un factor de Bayes de 19.4, las probabilidades a posteriori son sólo $p_{posterior} = 0.196$ y las probabilidades a priori, dada la baja tasa de base, son $p_{prior} = 0.0101$. Así pues, aunque los datos han provocado un cambio masivo de las expectativas en términos absolutos, no basta con suponer de repente que la enfermedad es especialmente probable en vista de un resultado positivo de la prueba. La enfermedad sigue siendo bastante improbable. Esto demuestra cómo los índices de base afectan a los cálculos. Sin embargo, debemos ser conscientes de que en el contexto bayesiano siempre existe incertidumbre con respecto a todos los parámetros: las tasas base, las tasas de detección y las tasas de falsas alarmas. La incertidumbre existe no sólo en relación con la enfermedad, sino también en relación con las características de la prueba o en relación con la prevalencia en la población. Todo esto no se tuvo en cuenta en el estudio de caso.

Un ejemplo de pensamiento muestra la fácil sobreinterpretabilidad de los factores de Bayes en un modelo lineal simple. Una vez se toma una Prior por defecto de `brm()` y la otra vez se elige una Prior ligeramente informada pero posiblemente completamente errónea. Probar ambos modelos entre sí puede dar como resultado un enorme factor de Bayes que, si se malinterpreta, sugiere que un modelo es increíblemente superior al otro. Sin embargo, los resultados *a posteriori* son prácticamente idénticos, es decir, los modelos y sus parámetros son prácticamente indistinguibles a nivel numérico a posteriori. ¿Qué dice entonces el enorme factor de Bayes? Sólo que los datos han cambiado masivamente las expectativas (la Prior), porque los factores de Bayes proporcionan información relativa sobre cómo un modelo comparado con otro conduce a un cambio en las expectativas dados los datos. El cociente de estos dos cambios en las expectativas entre los dos modelos, que *sólo* difieren en las expectativas previas, indica de nuevo que el cambio en las expectativas es mucho mayor en un modelo que en el otro. Si la Prior de un modelo es mucho más parecida a la Posterior que la otra Prior, porque una representa un ajuste informado y ya bueno a la Posterior y la otra un ajuste muy inespecífico, esto no es sorprendente. Entonces el modelo con la Prior inespecífica conduce a un enorme cambio en las expectativas dados los datos, mientras que esto no es de esperar con el modelo utilizando la Prior informada. Con conjuntos de datos suficientemente grandes, la Posterior debería estar determinada principalmente por los datos (Likelihood) y ya no por la Prior. En consecuencia, a pesar de un enorme factor de Bayes, la Posterior resulta ser casi idéntica dentro del ámbito de las diferencias habituales entre cadenas MCMC.

¿Debería basarse ahora una decisión entre modelos en un enfoque de este tipo, que utiliza sólo parte de la información y expresa únicamente un cambio relativo en las expectativas? En nuestra opinión, se debe tratar esto con cautela y supone que se han eliminado todas las incertidumbres sobre la elección de la Prior y que es esencial incluir los valores de la Posterior en la prueba y validación del modelo. A continuación, los factores de Bayes, así como los *p*-valores proporcionan una perspectiva adicional sobre los datos y los

modelos, pero sin determinar las decisiones. No se encuentra la significación basado solamente en las matemáticas, sino en toda la situación de la investigación.

He aquí un ejemplo de datos. Nos basamos en los datos de Gürtler (2005): datos de un cuestionario sobre el humor. Se comprobó que la cantidad de texto producida por los sujetos difería entre géneros y tipos de escuela (para más detalles sobre el ajuste del modelo, véase el capítulo 6.8.4.6). El resultado fue un modelo lineal que predecía la producción logarítmica de palabras por varios factores. Entre ellos, la edad en la escala logarítmica, edad dividida en tres categorías (véase Dalgaard, 2004), tipo de escuela y el sexo. Además, se permitieron diferentes varianzas para el tipo de escuela y el género. Para estimar el modelo de modo bayesiano utilizamos `brm()` del paquete R `brms` (`ptII_quan_Bayes_BayesFactors_further-remarks.r`).

```
# data LG word frequency schooltype age sex R-Code
# read data
diss <- read.csv("LG_school-words-raw.tab", header=TRUE, sep="\t")
str(diss)
head(diss)
namen <- names(diss)
diss$SS <- paste(diss$sex,diss$schooltype,sep="")
diss$stype <- factor(diss$stype)
diss$sex <- factor(diss$sex)
dim(diss)
# create categories for age with different breaks
diss$age.cat <- cut(as.numeric(diss$age), breaks=c(13,16,19,25),
               include.lowest=TRUE, right=TRUE,
               labels=c("(14-16]","(17-19]","(20-25]"))
diss$age.cat <- factor(diss$age.cat)
diss$age.cat1 <- cut(as.numeric(diss$age), breaks=c(13,19,25),
               include.lowest=TRUE, right=TRUE,
               labels=c("(14-19]","(20-25]"))
diss$age.cat1 <- factor(diss$age.cat1)
diss$W.noSC <- as.integer(diss$W.noSC)
diss$W.noSC.log <- log(diss$W.noSC)
diss$age.log <- log(diss$age)
diss$SS <- factor(diss$SS)
diss$schooltype <- factor(diss$schooltype)
# as.numeric(factor(paste(diss$sex,diss$schooltype,sep="")))
diss$SSn <- as.numeric(factor(diss$SS))
diss$sexn <- as.numeric(factor(diss$sex))
diss$stypen <- as.numeric(factor(diss$schooltype))
head(diss)
yName <- "W.noSC.log"
xName <- c("age.log","sexn","stypen")
fileNameRoot <- "Diss-"
numSavedSteps <- 15000
thinSteps <- 50
# get it from Kruschke...
source("DBDA2E-utilities.R")[1]
source("Jags-Ymet-XmetMulti-Mrobust.R")[2]
# prepare what we want to analyze and remove NAs
diss1 <- diss[,c(yName,xName)]
naid <- which(is.na(diss1), arr.ind=TRUE)[,1]
diss.nona <- diss1[-naid,]
```

Después de la estimación del modelo

[1] DBDA2E-utilities.R

[2] Jags-Ymet-XmetMulti-Mrobust.R

```
# t11
diss.res.t11 <- brm(bf(log(W.noSC) ~ log(age) + age.cat1 + stype + sex,
                sigma ~ 0 + stype * sex), data=diss,
                family=student(), save_all_pars=TRUE)
```

se calcula un segundo modelo que es idéntico al primero, salvo por las diferentes hipótesis a priori. Para crear verdaderas Priors informadas, sería necesario un estudio intensivo de la literatura pertinente. El estudio original de Gürtler (2005) utilizó métodos frecuentistas en lugar de bayesianos. Por lo tanto, no disponemos de un análisis de este tipo. Prescindimos de él y aplicamos una Prior ligeramente informada, a saber, N(0; 3) para los β-coeficientes y Cauchy(0; 3) para las varianzas.

```
# t11 diff priors
# just a model
model.stan.t11 <- bf(log(W.noSC) ~ log(age) + age.cat1 + stype + sex,
sigma ~ 0 + stype * sex)
model.stan.t11
prior.t11 <- get_prior(model.stan.t11, data=diss)
# define our own priors, no theoretical background intended...
prior.t11.pr1 <- c(
prior(normal(0,3), class=b),
    prior(normal(0,3), class=b, coef="age.cat12OM25"),
    prior(normal(0,3), class=b, coef="logage"),
    prior(normal(0,3), class=b, coef="sexw"),
    prior(normal(0,3), class=b, coef="stypeG"),
    prior(normal(0,3), class=b, coef="stypeR"),
    prior(student_t(3,5,10), class=Intercept),
    prior(cauchy(0,3), class=b, dpar=sigma),
    prior(cauchy(0,3), class=b, coef="sexw", dpar=sigma),
    prior(cauchy(0,3), class=b, coef="stypeB", dpar=sigma),
    prior(cauchy(0,3), class=b, coef="stypeG", dpar=sigma),
    prior(cauchy(0,3), class=b, coef="stypeG:sexw", dpar=sigma),
    prior(cauchy(0,3), class=b, coef="stypeR", dpar=sigma),
    prior(cauchy(0,3), class=b, coef="stypeR:sexw", dpar=sigma)
)
prior.t11.pr1
identical(prior.t11,prior.t11.pr1)
```

Esta distribución a priori se pasa a la llamada de `brm()`.

```
diss.res.t11.pr1 <- brm(bf(log(W.noSC) ~ log(age) + age.cat1 + R-Code
stype + sex, sigma ~ 0 + stype * sex),
    data=diss, family=student(),
    prior=prior.t11.pr1, save_all_pars=TRUE)
prior_summary(diss.res.t11, data=diss)
prior_summary(diss.res.t11.pr1, data=diss)
prior.t11.pr1
```

A continuación, se utiliza `bayesfactor()` para contrastar los dos modelos.

```
# nonsense! almost same results...
# but different update of the prior knowledge!
# be aware that this is enough to change
# a Bayes Factor EVEN if the model basically is identical!
# and even estimations are basically identical...
BF <- bayes_factor(diss.res.t11, diss.res.t11.pr1)
BF # clearly PRO 'diss.res.t11'
```

El factor de Bayes resultante de 111 571 984 131.60162 (≈ 111 mil millones) es exorbitantemente grande para el modelo con la Prior por defecto. Como recordatorio – los cálculos pueden diferir para los lectores porque no hemos pasado a `brm()` un valor inicial `seed` para el generador aleatorio. Podemos hacer que estos resultados sean más manejables por expresarlos con múltiplos de potencias de diez:

```
> BF # clearly PRO 'diss.res.t11'
Estimated Bayes factor in favor of
diss.res.t11 over diss.res.t11.pr1: 113440056702.45285
> format(BF$bf, scientific=TRUE)
[1] "1.134401e+11"
```

En cambio, los propios parámetros posteriores son prácticamente idénticos para los β-coeficientes respectivos. Observamos las desviaciones en porcentaje:

```
> # re-check coefficients percent
> round((1 - fixef(diss.res.t11) / fixef(diss.res.t11.pr1))*100,3)
                   Estimate Est.Error Q2.5    Q97.5
Intercept          -0.010   -3.500    7.211   -0.687
logage             -0.073   -3.478   -3.665   -4.745
age.cat12OM25      -1.424   -0.766   -3.078   40.847
stypeG              1.215   -1.348    9.316    0.772
stypeR             -1.950   -0.950   -4.066    1.820
sexw               -0.189    1.584   -0.615    0.165
sigma_stypeB       -1.099    1.760   -0.711   -2.366
sigma_stypeG       -0.229    1.986    1.572 -139.613
sigma_stypeR       -0.215   -1.568   -0.638   -0.128
sigma_sexw        -35.253  -14.778   15.979  -21.824
sigma_stypeG:sexw -11.632  -13.127  -12.624  -18.546
sigma_stypeR:sexw -46.876  -14.987  -24.406   14.314
```

Se producen desviaciones para las varianzas del género y las de la interacción del tipo de escuela y el género. Surgen otras desviaciones para los límites del cuantil Q97.5 y, condicionalmente, para el cuantil inferior Q2.5. Pero para estos parámetros las estimaciones reales de los parámetros son las mismas en términos de dirección, pero difieren en cierta medida en términos de magnitud y, como se ha mencionado, en los intervalos de confianza – ninguna de las cuales es realmente sustancial, por lo que saldrían a la luz nuevas conclusiones. La influencia de la Prior en el cálculo de las varianzas es mencionada por Kruschke (2015a) y por lo tanto es atribuida por nosotros a la elección de las diferentes Priors y no a la validez del modelo per se. Con la Prior ligeramente informada, los errores estándar de los parámetros relacionados con la varianza son algo menores porque no permitimos una Prior plana. Con una inversión mucho mayor a nivel de contenido, los valores de la Prior podrían mejorarse y reducirse considerablemente. Veamos el coeficiente bayesiano de determinación con `bayes_R2()`, parece prácticamente idéntico:

```
> bayes_R2(diss.res.t11)
    Estimate  Est.Error  Q2.5      Q97.5
R2  0.3827624 0.02596181 0.3301151 0.4314158
> bayes_R2(diss.res.t11.pr1)
    Estimate  Est.Error  Q2.5      Q97.5
R2 0.3830838  0.02609619 0.3305646 0.4330482
```

Si pasamos al nivel de comparación de modelos, una comprobación gráfica de predicción posterior utilizando `pp_check()` no muestra diferencias significativas entre los dos modelos. Una comparación numérica utilizando los criterios de información `LOO()` y `WAIC()` del paquete R `loo` y `loo_R2()` de `brms` no revela diferencias realmente significativas entre los modelos. A pesar de las Priors claramente diferentes, los resultados numéricos son prácticamente *idénticos*.

```
# information criteria
LOO1 <- LOO(diss.res.t11, reloo=TRUE)
LOO2 <- LOO(diss.res.t11.pr1, reloo=TRUE)
WAIC1 <- WAIC(diss.res.t11)
WAIC2 <- WAIC(diss.res.t11.pr1)
```

Ignoramos los mensajes de advertencia. En un estudio serio, tendríamos que mirar más de cerca los datos 317 comunicados.

```
> loo_compare(x=list(LOO1, LOO2))
                   elpd_diff se_diff
diss.res.t11.pr1     0.0       0.0
diss.res.t11        -1.2       1.0
> loo_compare(x=list(WAIC1, WAIC2))
                   elpd_diff se_diff
diss.res.t11.pr1     0.0       0.0
diss.res.t11        -0.1       0.3
```

La diferencia entre los dos modelos es insignificante. El procedimiento podría continuarse con un `gráfico()` de las distribuciones posteriores o de los efectos con `marginal_effects()`. De este modo, los criterios de información y los análisis gráficos (predictivos) contrastan con el factor de Bayes anterior: al fin y al cabo, un número enorme de varios miles de millones, pero despreciable a posteriori.

```
# refit models for LOO
loo1 <- loo.brmsfit(diss.res.t11, reloo=TRUE)
loo2 <- loo.brmsfit(diss.res.t11.pr1, reloo=TRUE)
loo1
loo2
```

Comparamos la salida:

```
> # compare
> loo_compare(x=list(loo1, loo2))
                   elpd_diff se_diff
diss.res.t11.pr1    0.0       0.0
diss.res.t11       -1.1       0.7
```

y la representamos gráficamente (véase la Fig. 6.29, dis.res.t11 arriba y dis.res.t11.pr1 abajo):

```
# plot LOO R-Code
plot(loo1)
plot(loo2)
```

y procedeos al coeficiente de determinación corregido por `loo`. Ignoramos deliberadamente el el diagnóstico `Pareto-k`. En la práctica, por supuesto, habría que examinarlo más detenidamente, ya que indica que el modelo aún no es óptimo:

```
> # information criteria, R2 corrected for LOO
> brms::loo_R2(diss.res.t11)
    Estimate Est.Error Q2.5  Q97.5
R2  0.357    0.0404    0.275 0.433
Warnmeldung:
Some Pareto k diagnostic values are too high.
See help('pareto-k-diagnostic') for details.
> brms::loo_R2(diss.res.t11.pr1)
```

```
    Estimate Est.Error Q2.5  Q97.5
R2  0.355    0.041     0.273 0.431
Warnmeldung:
Some Pareto k diagnostic values are too high.
See help('pareto-k-diagnostic') for details.
```

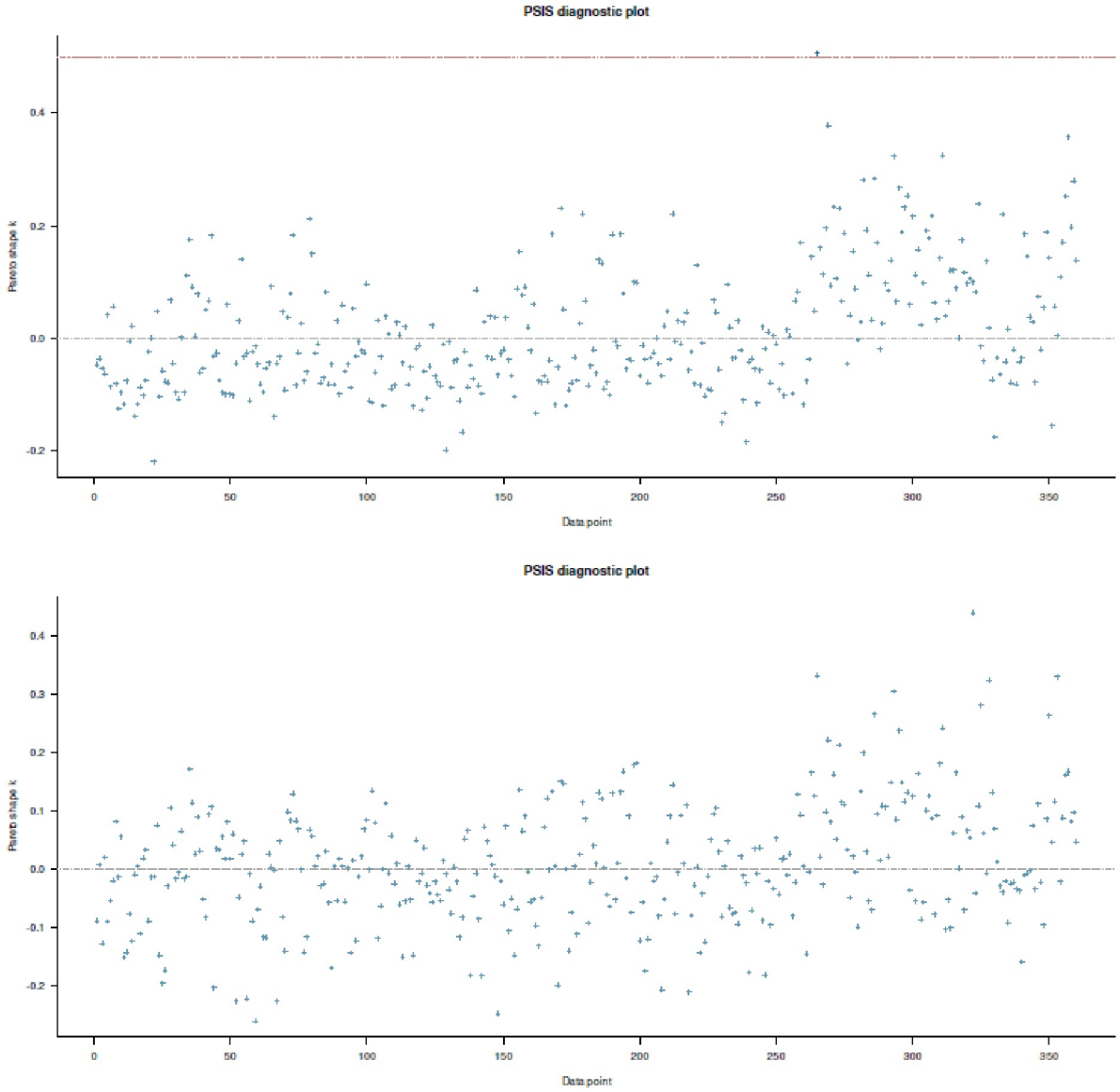


Figura 6.29. *Estudio de Gürtler (2005, comparación de modelos, 1oo)*

Vemos las Posteriors así (véanse las Figs. 6.30, 6.31 y 6.32, cada vez dis.res.t11 arriba y dis.res.t11.pr1 abajo):

```
# plot posteriors
plot(diss.res.t11)
plot(diss.res.t11.pr1)
```

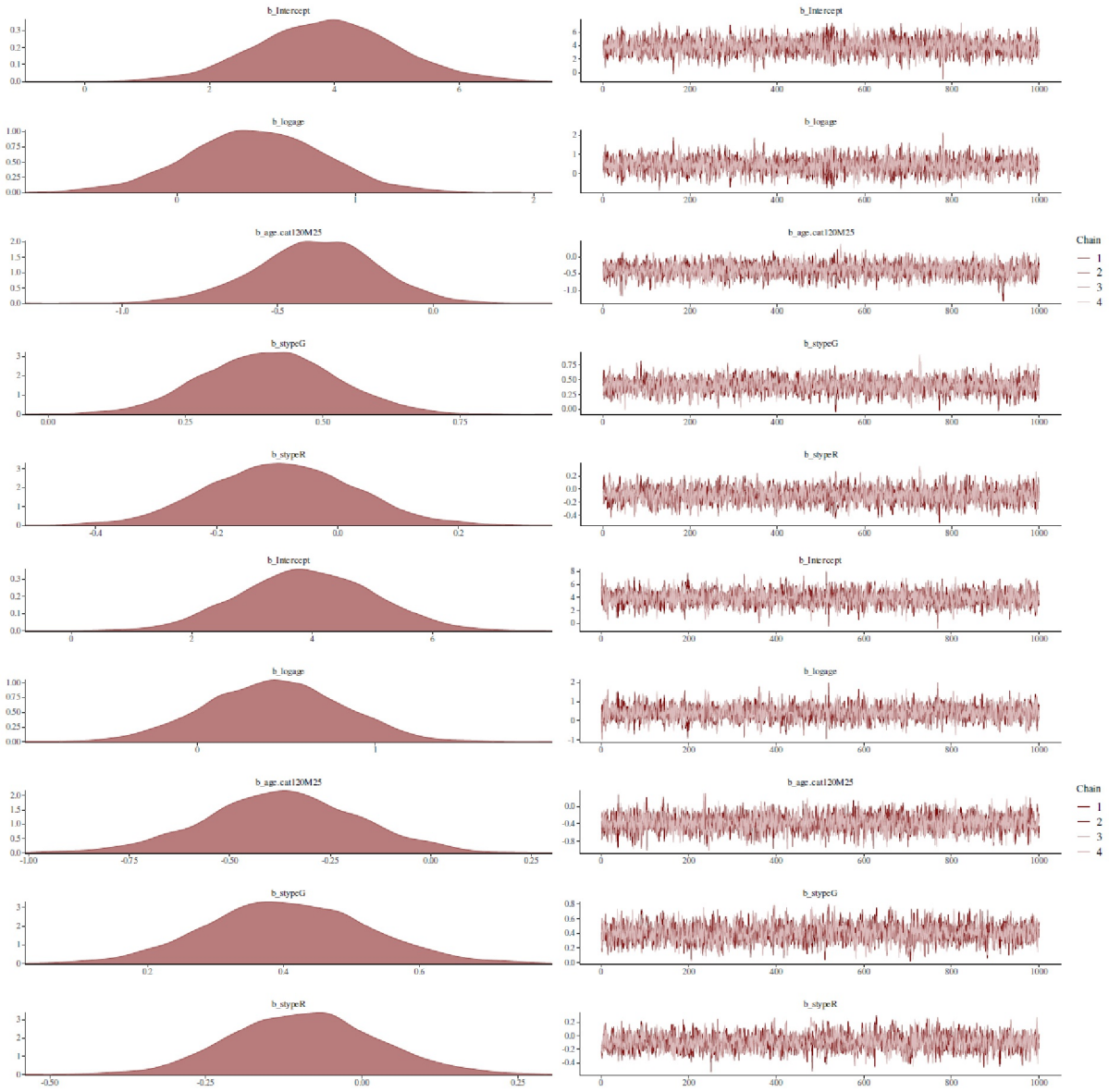


Figura 6.30. *Estudio de Gürtler (2005, comparación de modelos, Posterior, parte 1)*

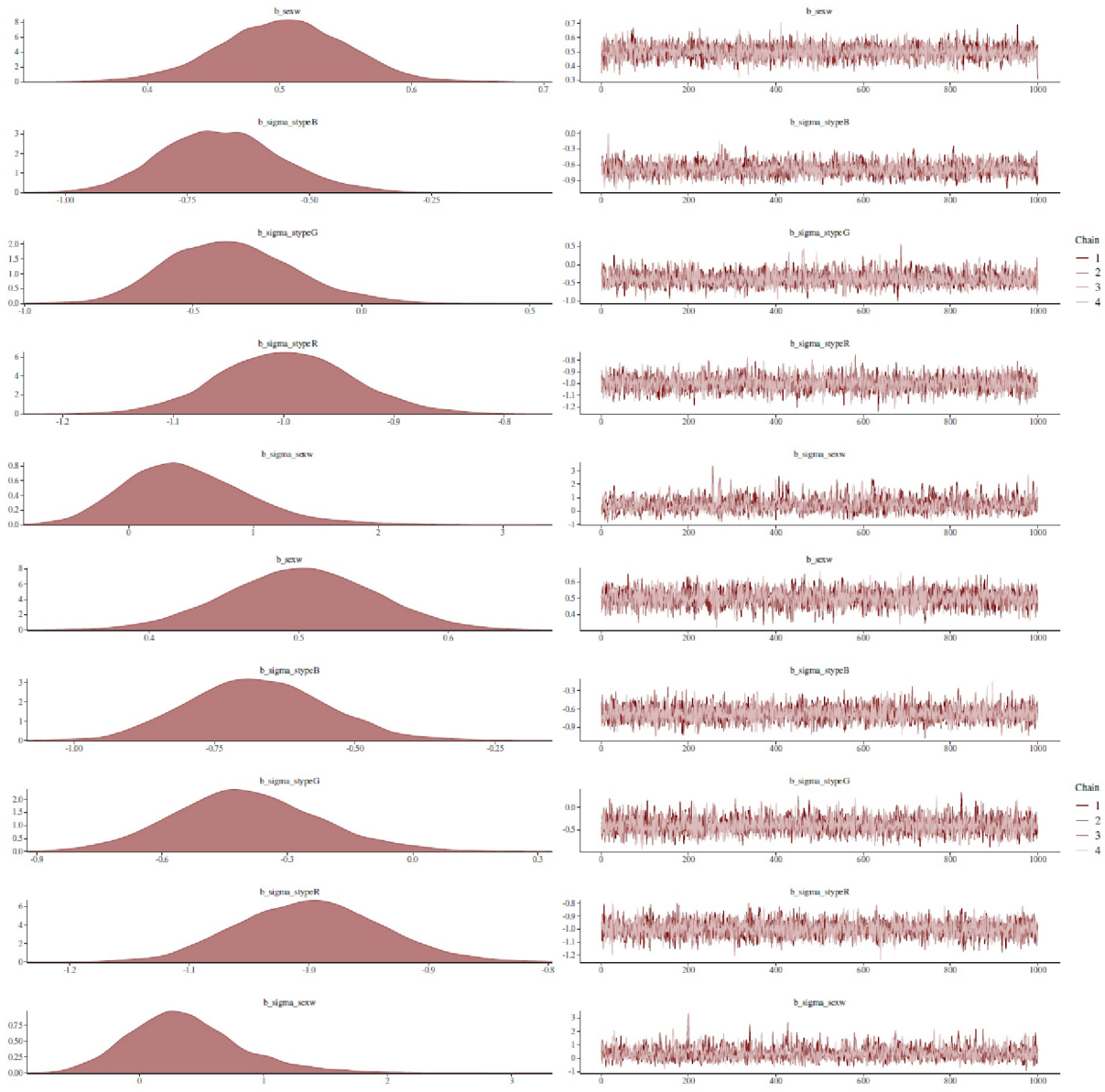


Figura 6.31. *Estudio de Gürtler (2005, comparación de modelos, Posterior, parte 2)*

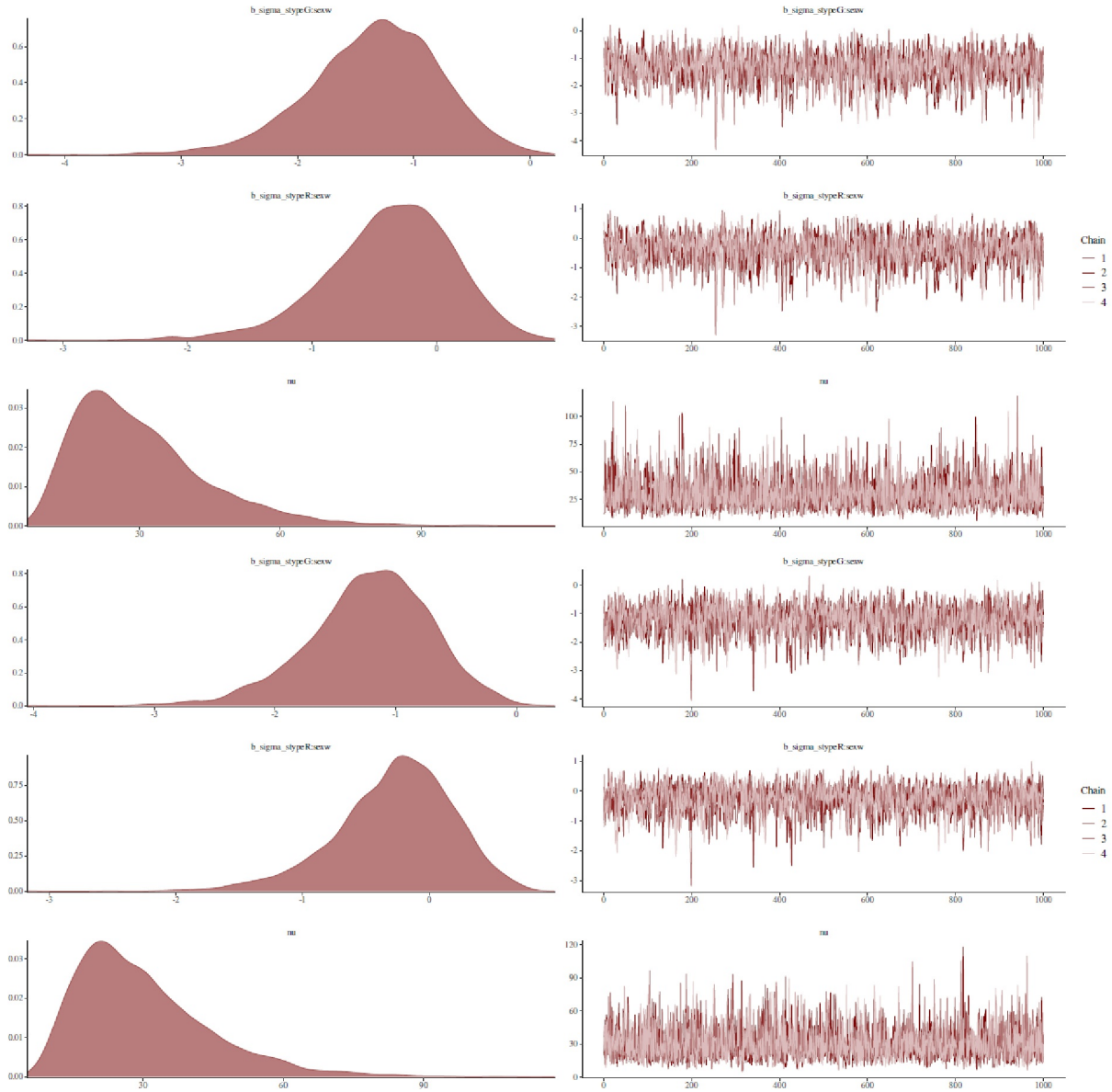


Figura 6.32. *Estudio de Gürtler (2005, comparación de modelos, Posterior, parte 3)*

Los efectos tienen este aspecto (véase la Fig. 6.33 diss.res.t11 a la izquierda y diss.res.t11.pr1 a la derecha):

```
# plot marginal effects
marginal_effects(diss.res.t11)
marginal_effects(diss.res.t11.pr1)
```

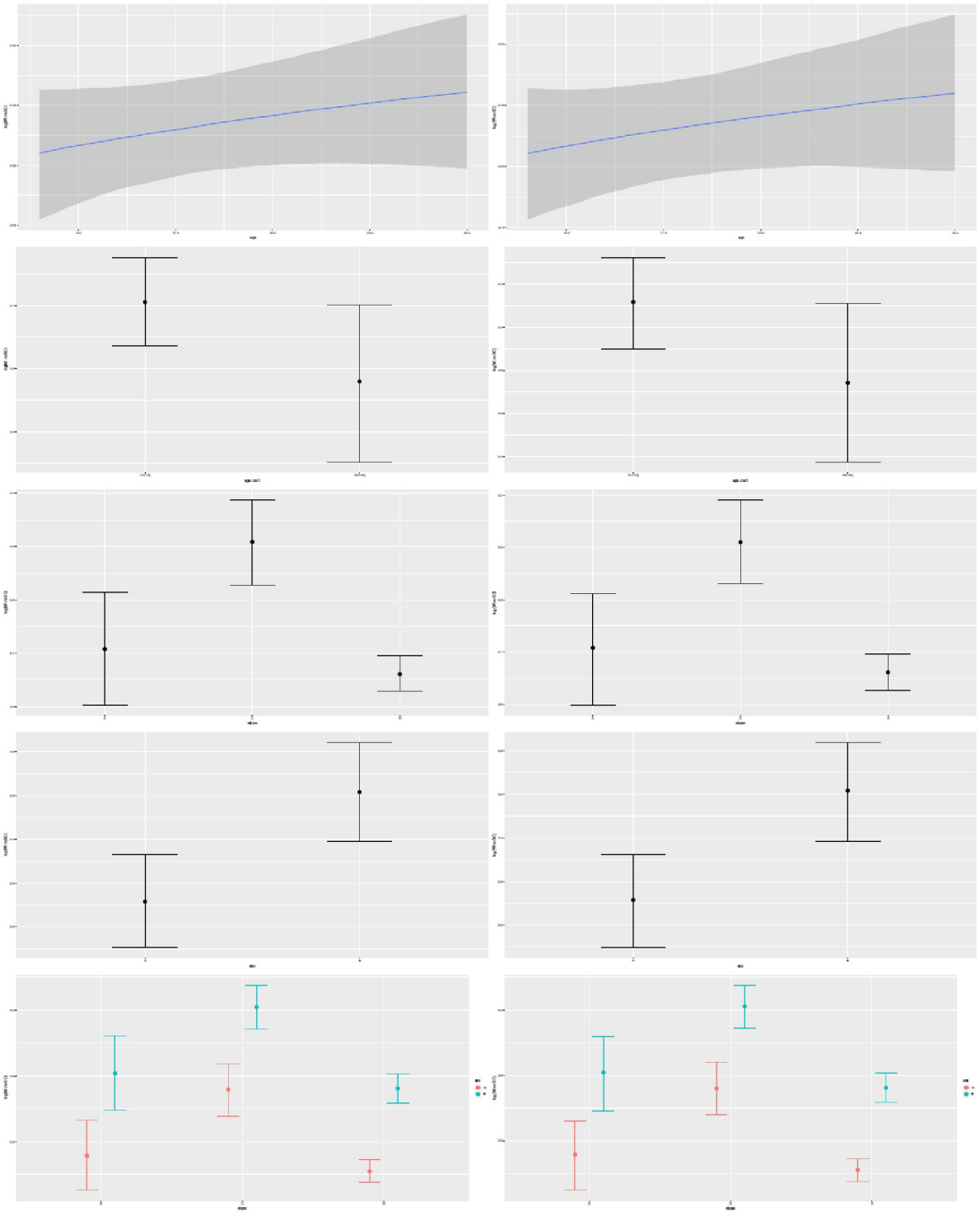


Figura 6.33. *Estudio de Gürtler (2005, comparación de modelos, efectos)*

Una vez más, el factor de Bayes expresa un cambio en las expectativas y no sustituye en modo alguno a una evaluación exhaustiva o incluso absoluta del modelo. Sin embargo, para comparar modelos, debe incluirse toda la información disponible, para lo cual los factores de Bayes por sí solos parecen inadecuados. La situación es diferente si se incluye un nuevo factor en el modelo sobre una base de prueba y ensayo con

una Prior idéntica (véase también Kruschke, 2013c, sobre el tema de la evaluación de modelos). En este caso, un factor Bayes puede contribuir a la medida en que el factor conduce a un cambio significativo en la expectativa a priori. No obstante, esto no obvia la necesidad de observar detenidamente la Posterior y de examinar los parámetros en cuanto a su plausibilidad (dirección y tamaño del efecto, véase el capítulo 4.3.3.2), así como de evaluar la calidad predictiva del modelo gráfica y numéricamente. Además, se realiza una reflexión teórica crítica de los efectos.

Así pues, cuando se trata de comparar modelos, merece la pena echar un vistazo a los criterios de información comunes que, aunque tienen nombres bayesianos, se utilizan de forma interesante en la estadística frecuentista.

6.8.2 Criterios de información

En la actualidad, los denominados *criterios de información* se utilizan en la estadística frecuentista como alternativa a los valores *p* para examinar la superioridad relativa de los modelos entre sí. De ello se pueden derivar decisiones a favor de un determinado modelo. Dziak, Coman, Lanza y Li (2012) y Watanabe (2013) ofrecen una visión general de los criterios de información más comunes y los discuten con sus ventajas e inconvenientes. Los valores característicos comunes son el criterio de información bayesiano *BIC* (Schwarz, 1978), el criterio de información de Akaike *AIC* (Akaike, 1983) y diversos valores derivados y ajustados, es decir, corregidos, como *BICc* y *AICc* (Bozdogan, 2000), respectivamente, así como el DIC (criterio de información de desviación, Spiegelhalter, Best, Carlin & van der Linde, 2014), que está estrechamente relacionado con el *AIC*. Las soluciones ajustadas tienen en común que cada una de ellas calcula el correspondiente criterio de goodness of fitting (bondad de ajuste) *GOF* y penaliza un aumento constante de los parámetros del modelo (= *sobreajuste*), lo que conduce a una gradación del valor del criterio de información. En general, los criterios *GOF* examinan lo bien que un modelo estadístico describe los datos empíricos. Normalmente, la diferencia entre el modelo (= expectativa) y los datos empíricos (= observación) entra en la ecuación, lo que corresponde a un análisis residual en sentido amplio. A continuación, se puede examinar si los datos siguen una determinada distribución, por ejemplo, la cuestión de la distribución normal de los residuos. Además de la diferencia l, que se estima mediante la máxima Likelihood en su forma logaritmizada ln(^ L), el tamaño de la muestra N y el número de parámetros estimados del modelo k entran en los criterios de información enumerados aquí.

La Likelihood no es una probabilidad, ya que puede asumir valores superiores a 1 debido a un factor. Sin embargo, es proporcional a una probabilidad. La probabilidad es una función de las posibles realizaciones de los datos en vista de los parámetros del modelo, mientras que la Likelihood cubre el caso inverso, es decir ella es una función de las posibles realizaciones de los parámetros del modelo en vista de los datos. El objetivo de la probabilidad es describir o predecir la posibilidad de que se produzca una situación concreta. La probabilidad se utiliza para maximizar las posibilidades de que se produzca una situación concreta. Conceptualmente, la Likelihood tampoco es una probabilidad, ya que en la estadística frecuentista los parámetros se entienden como una cantidad desconocida pero fija. No se manejan como variables aleatorias. Para ilustrar esto, podemos imaginar que a nosotros nos gustaría saber la edad de una cierta persona, pero no podemos preguntar por razones no especificadas. La edad es desconocida, aunque podemos sacar ciertas conclusiones a partir de cierta información sobre la persona (aspecto, lenguaje, profesión, comportamiento, etc.). Esto significa que podemos restringir la edad a un cierto límite de tolerancia en el sentido de, por ejemplo "Hay una probabilidad de 0,4 de que la edad sea 37 años". Sin embargo, esta estimación no convierte la edad en una variable aleatoria que tenga un 40% de probabilidades de ser 37 años. La edad es simplemente una cantidad desconocida, pero fija y aquí continuamente cambiante. Esta afirmación sería posible sin ningún problema en el marco de la estadística bayesiana. Con respecto a la PDF (= función de densidad de probabilidad), se maneja de tal manera que con la Likelihood la observación es fija y los valores de los parámetros están distribuidos, mientras que con la PDF de una probabilidad el parámetro se mantiene fijo y las observaciones están distribuidas. Mientras que la suma de las distintas observaciones en la PDF de una probabilidad es 1, no ocurre lo mismo con la Likelihood si se suman todos los valores de los parámetros en una observación fija. La suma puede ser incluso infinitamente grande. La

Likelihood forma parte del teorema de Bayes Posterior = PriorLikelihood / TotalProbability, de modo que, desde este punto de vista, la estadística frecuentista como parte del teorema de Bayes está subordinada a él en cierto modo.

La base de todas las restricciones es la *log-Likelihood*. Del logaritmo natural de la función de Likelihood *L()*.

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} f_i(y_i \mid \theta) \tag{6.57}$$

se calcula

$$\ln \mathcal{L}(\theta) = \sum_{i=1}^{N} \ln(f_i(y_i \mid \theta)) \tag{6.58}$$

La estimación de la máxima Likelihood resulta como

$$\hat{\mathcal{L}}(\theta) = \arg\max \mathcal{L}(\theta) \tag{6.59}$$

Usualmente se trata de maximizar la forma logmarítmica:

$$\ell = \ln(\hat{\mathcal{L}}(\theta)) \tag{6.60}$$

Es importante comprender que no se trata de soluciones bayesianas completas que requieran una selección cuidadosa de la Prior. Más bien, una Prior plana, es decir, una distribución uniforme, no sólo se suele suponer implícitamente, sino que se aplica directamente (McElreath, 2015). La aplicación de criterios de información en contextos clásicos descuida una selección razonada de la Prior. Por lo tanto, nunca es una solución bayesiana completa.

El BIC se basa en la idea de que para muchos modelos el BF_{ij} de dos modelos *i* frente a *j* viene dado por la expresión $exp((BIC_i - BIC_j)/2)$ (Wagenmakers, 2007a, p.796). El Raftery (1999) y Bollen, Surajit, Zavisca y Harden (2011) analizan la relación entre BIC y BF.

El BIC y el AIC se calculan del siguiente modo (Bozdogan, 2000):

$$\mathrm{BIC} = -2 \cdot \ell + \ln(N) \cdot k \tag{6.61}$$

o en la versión modificada BIC_c con b = bias

$$\mathrm{BIC}_c = -2 \cdot \ell + 2 \cdot N \cdot b \tag{6.62}$$

El AIC se calcula de forma algo diferente:

$$\mathrm{AIC} = -2 \cdot \ell + 2 \cdot k \tag{6.63}$$

y en la versión corregida AIC_c

$$\mathrm{AIC}_c = n \cdot \log(2 \cdot \pi) + n \cdot \log(\hat{\sigma}^2) + n + 2 \cdot \frac{n \cdot (k+1)}{n-k-2} \tag{6.64}$$

Las versiones exactas corregidas AIC_c y BIC_c se establecen a lo largo del modelo estadístico (Bozdogan, 2000). Por lo tanto, no existe una forma general sencilla.

Se cumple que AIC_c para$N \to \infty$ converge a AIC y el factor de penalización tiende a cero. Se considera que el BIC es consistente, de modo que a medida que crece la muestra, selecciona el modelo óptimo, es decir, el que explica el máximo con el menor número de parámetros. Según Schwarz (1978), el BIC considera que tanto los errores de tipo I como los de tipo II (véase el capítulo 4.3.3) son igualmente indeseables, mientras que otros criterios, como el AIC, consideran que los errores de tipo II son más indeseables que los de tipo I, a menos que el tamaño de la muestra sea muy pequeño (véase la discusión con un caso práctico ficticio sobre la elección de la dirección de la hipótesis en relación con los errores de tipo I y de tipo II en el capítulo 4.3.3.4).

El DIC se determina mediante la desviación $D(\theta)$ y el número efectivo de parámetros del modelo p_D, este último estimado mediante la diferencia de $\bar{D}$ y $\hat{D}$ en la ubicación de la media posterior. Si θ representa los parámetros desconocidos, $\bar{\theta}$ como el valor esperado de θ (= media posterior), y son los datos, $p(y|\theta)$ es la función de Likelihood y C es una constante que siempre se trunca mutuamente en la comparación de modelos, entonces se aplica lo siguiente:

$$D(\theta) = -2 \cdot \log(p(y|\theta)) + C \tag{6.65}$$

Para p_D, hay las versiones según Spiegelhalter, Best, Carlin y van der Linde (2002) o Gelman, Carlin, Stern y Rubin (2004, p.182):

$$\begin{aligned} p_{D\,(\text{Spiegelhalter et al.})} &= \overline{D} - D(\overline{\theta}) \\ &= \overline{D} - \hat{D} \end{aligned} \tag{6.66}$$

$$p_{D\,\text{Gelman et al.}} = \frac{1}{2} \cdot \widehat{var}(D(\overline{\theta})) \tag{6.67}$$

El resultado es que el DIC es la diferencia entre la desviación media de cada parámetro en la Posterior D y el número de parámetros efectivos del modelo p_D,

$$\begin{aligned} \text{DIC}_{\text{Spiegelhalter et al.}} &= p_D + \overline{D} \\ &= D(\overline{\theta}) + 2 \cdot p_D \end{aligned} \tag{6.68}$$

$$\begin{aligned} \text{DIC}_{\text{Gelman et al.}} &= \overline{D} + (\overline{D} - \hat{D}) \\ &= \overline{D} + p_D \end{aligned} \tag{6.69}$$

Las fórmulas anteriores muestran que no se incluyen Priores basados en el contenido de los criterios de información. El modelo con el AIC, BIC, DIC, etc. más bajo corresponde al que tiene el factor de Bayes más alto. Recuerde que un factor de Bayes es la ratio de Likelihood de dos Marginal Likelihoods, cada una de las cuales representa dos hipótesis contrapuestas. No depende de un conjunto concreto de parámetros porque todos ellos se han integrado. La integración hace que todo el rango de valores de los parámetros se evalúa (es decir, se tiene en cuenta), se pondera por la Prior y, a continuación, se suma. Si y son los datos, M es el modelo y θ los parámetros del modelo, se aplica esta integral a la Marginal Likelihood:

$$p(y|M) = \int p(y|\theta, M) \cdot p(y|\theta, M)\, d\theta \tag{6.70}$$

Veamos el código R correspondiente. Primero definimos una función del producto de la Prior y la Likelihood, que vamos a integrar (`ptII_quan_Bayes_infocriteria.r`):

```
plik <- function(theta) {
dbinom(x=40, size=150, prob=theta) *
dbeta(x=theta, shape1=3, shape2=1)
}
```

Sigue la integración con `integrate()`:

```
margLik <- integrate(f=plik, lower=0, upper=1)$value
margLik
```

Si se sustituye la integral del factor de Bayes por la estimación de máxima Likelihood, se obtiene la prueba clásica deL ratio de Likelihood. Si las dos hipótesis tienen las mismas Priores, el factor de Bayes corresponde al Posterior Odds-Ratio.

Los criterios de información bayesianos, como WAIC (= Widely Applicable Information Criteria; Vehtari, Gelman & Gabry, 2017; Gelman, Hwang & Vehtari, 2013) o como WAIC (= Watanable-Akaike, véase *LOO* = *Leave-One-Out Cross Validation*,Watanabe, 2010, generalizable a *Leave-k-Out Cross Validation*) implican la solución completa del teorema de Bayes mediante el cálculo de la Likelihood (por razones numéricas, como es *habitual* en forma logarítmica) sobre toda la Posterior. WAIC determina la incertidumbre de las predicciones con una precisión milimétrica, es decir, caso por caso a partir de los datos. Esto tiene la ventaja de que los datos con diferente predictibilidad se incluyen en su diversidad, cada uno como observacion independiente. Estas comparaciones individuales se suman en todas las comparaciones. Si $p(y_i)$ es la probabilidad promediada de la observación *i* en el conjunto de datos de entrenamiento, entonces WAIC resulta de la densidad de predicción puntual logaritmizada *lppd* (= *log-pointwise-predictive-density*)

$$\text{lppd} = \sum_{i=1}^{N} \log(p(y_i)) \tag{6.71}$$

y el número efectivo de parámetros p_{WAIC}. Si $V_{(yi)}$ es la varianza de la log-Likelihood de la observación *i* en la muestra de entrenamiento, que resulta de la Posterior, p_{WAIC} se denota por

$$\text{p}_{\text{WAIC}} = \sum_{i=1}^{N} V(y_i) \tag{6.72}$$

y resulta WAIC como

$$\text{WAIC} = -2 \cdot (\text{lppd} - p_{\text{WAIC}}) \tag{6.73}$$

McElreath (2015, p.192) señala: "Y este valor es otra estimación de la desviación fuera de la muestra". Así pues, el *lppd* corresponde a un valor exacto de desviación, promediado sobre la Posterior. Por el contrario, la medida de desviación *DIC* introducida anteriormente, al igual que *SIC*, requiere una Posterior gaussiana, que en el caso de las desviaciones (por ejemplo, Posterior sesgada) puede conducir a resultados no interpretables. Con una Priori uniforme plana, *DIC* se reduce a *AIC* . En R, *WAIC* o *LOO* pueden calcularse mediante `LOO()` y `WAIC()` del paquete R `brms` o con `WAIC()` del paquete R `rethinking`. Este último también ofrece la medida de desviación *DIC* mediante `DIC()`. El paquete R `loo` está dedicado a la validación cruzada mediante la omisión selectiva de datos a través de `loo()`, `waic()` y otras medidas de

comparación de modelos. El paquete R `brms` hace lo mismo. McElreath (2015) muestra cómo se puede calcular manualmente *WAIC* a partir de la Posterior. Según el autor, una crítica de *WAIC* es el requisito de observaciones independientes, que no puede cumplirse en el caso de mediciones repetidas y series temporales. Esto no impide la salida numérica de *WAIC*, pero deja abierto qué significa entonces esta medida.

Por último, Watanabe (2013) añade otro criterio de información, esta vez denominado *WBIC* (= *Widely applicable Bayesian Information Criterion*). Se define como

$$\mathrm{WBIC} = \mathbb{E}_w^{\beta}[n \cdot L_n(w)]$$

con el parámetro β

$$\beta = \frac{1}{\log(n)}$$

El término $E^{\beta}_{w}[: : :]$ denota el valor esperado sobre la Posterior de *W* denotada como una función integrable arbitraria *G(w)* (ibid., p.869, fórmula 5). Se define β como $\beta > 0$ y eso representa la temperatura inversa y *n* representa el número de muestras de entrenamiento. La temperatura inversa óptima β* resulta de la convergencia de $\beta * \log(n) \rightarrow 1$ para $n \rightarrow \infty$. Según el autor, *WBIC* es aplicable cuando no se conoce información sobre la distribución verdadera. El *WBIC* amplía el *BIC* como versión generalizada en el contexto de los modelos singulares. Dado que el *WBIC* se desarrolló en el contexto del *machine learning*, un modelo estadístico es naturalmente singular si se desarrolló para extraer estructuras ocultas de un fenómeno arbitrario y aleatorio (ibíd.). Es decir, la asignación de un parámetro a una distribución de probabilidad no es inequívoca en el caso singular y, por tanto, ciertos problemas no son realmente resolubles. En el caso de una matriz, no tiene determinante alejado de cero y no es invertible. En relación con la matriz de información de Fisher, ésta ya no es definida positiva (= regular) en el caso singular. Resumido según Watanabe (ibíd., p.867),

> „In general, if a statistical model contains hierarchical layers, hidden variables, or grammatical rules, then it is singular. In other words, if a statistical model is devised so that it extracts hidden structure from a random phenomenon, then it naturally becomes singular. If a statistical model is singular, then the likelihood function cannot be approximated by any normal distribution, resulting that neither AIC, BIC, nor MDLcan be used in statistical model evaluation. Hence constructing singular learning theory is an important issue in both statistics and learning theory."

Con una pequeña función de R `exIC()`, se pueden obtener los criterios de información más comunes, como AIC, BIC, etc., para modelos lineales y modelos lineales (jerárquicos) generales. La función recopila funciones de diferentes paquetes de R que extraen dichos criterios de información y los agrupa. Una vez más, no existe el mejor criterio absoluto, ni bayesiano ni clásico, sino que cada uno tiene su idiosincrasia (véase también Dziak, Coman, Lanza & Li, 2012; McElreath, 2015, cap. 6). Veamos ejemplos en R (`ptII_quan_Bayes_information-criteria.r`).

`?glm` es un ejemplo de cómo, con n = 9 observaciones y 2 factores con 2 + 3 = 5 niveles de factor en realidad resulta en una gradación sustancial en AIC_c comparado con AIC. Con una relación más favorable entre los parámetros del modelo y el tamaño de la muestra, esta diferencia vuelve a desaparecer.

```
# Dobson (1990) Page 93: Randomized Controlled Trial
dobson.D93 <- data.frame(counts=c(18,17,15,20,10,20,25,13,12),
                outcome=gl(3,1,9),
                treatment=gl(3,3)
)
dobson.D93
glm.D93 <- glm(counts ~ outcome + treatment, family=poisson(), data=dobson.D93)
```

Antes de ver las diferencias, la situación es bastante diferente con otro ejemplo de conjunto de datos del mismo autor. Lo tomamos de la página de ayuda de R sobre `?lm`. Los criterios de información en cuestión difieren sólo modestamente entre sí, como se verá.

```
# lm
ctl <- c(4.17,5.58,5.18,6.11,4.50,4.61,5.17,4.53,5.33,5.14)
trt <- c(4.81,4.17,4.41,3.59,5.87,3.83,6.03,4.89,4.32,4.69)
dobson.D9 <- data.frame(weight=c(ctl,trt),
              group=gl(2, 10, 20, labels = c("Ctl","Trt")))
dobson.D9
lm.D9 <- lm(weight ~ group, data=dobson.D9)
```

Aquí la salida para ambos modelos:

```
> # example of huge penalty with AICcc ...
> # n=9 observations
> # k=5 (= 2 factors with 2 + 3 categories)
> glm.D93.exIC <- exIC(glm.D93)
fitting null model for pseudo-r2
Information Criteria for model type = glm
model = glm
N = 9
k = 5
df.resid = 4
AIC = 56.76
AICc = 76.76
BIC = 57.75
DIC = 5.129
lLH = -23.38
HQC = 54.63
crit Chi^2 = 9.488
Pearson Chi^2 = 5.173
phi = 1.293
sqrt(phi) = 1.137
llh [pR2s] = -23.38
llhNull [pR2s] = -26.11
G2 [pR2s] = 5.452
McFadden [pR2s] = 0.1044
r2ML [pR2s] = 0.4544
r2CU [pR2s] = 0.4557
> # less penalty
> lm.D9.exIC <- exIC(lm.D9)
fitting null model for pseudo-r2
Information Criteria for model type = lm
model = lm
N = 20
k = 2
df.resid = 18
AIC = 46.18
AICc = 46.88
BIC = 49.16
DIC = 8.729
lLH = -20.09
HQC = 44.57
crit Chi^2 = 28.87
Pearson Chi^2 = 8.729
phi = 0.485
sqrt(phi) = 0.6964
llh [pR2s] = -20.09
llhNull [pR2s] = -20.85
G2 [pR2s] = 1.518
McFadden [pR2s] = 0.0364
```

```
r2ML [pR2s] = 0.07308
r2CU [pR2s] = 0.08345
```

Para modelos lineales jerárquicos o generales, se sigue un procedimiento equivalente (ejemplos de las páginas de ayuda de R para `?lmer` y `?glmer`).

```
# lmer
fm1 <- lmer(Reaction ~ Days + (Days | Subject), data=sleepstudy)
fm1.exIC <- exIC(fm1)
# glmer
gm1 <- glmer(cbind(incidence, size-incidence) ~ period + (1|herd),
data=cbpp, family=binomial)
gm1.exIC <- exIC(gm1)
```

La función `exIC()` de R puede ampliarse fácilmente consultando el código fuente. Para eso está desarrollado. No es una función genérica universal de R que funcione como una biblioteca en todas partes. Por eso faltan algunos resultados para modelos lineales jerárquicos, por ejemplo, el del R-paquete `pscl`. Para ello, el esfuerzo de desarrollo sería bastante elevado para adaptar todo esto. En su lugar, representa una función generada ad-hoc que es útil en muchas situaciones para uso doméstico. Interesante, porque claramente expuesta, es la salida generada, que en primer lugar comienza con `RESULTADOS <- estructura(lista(...))` almacena los resultados en una lista y luego con

```
cat(paste(format(names(res), width = 17L, justify = "right"),
        format(res, digits = digits, nsmall=2), sep = " = "),
        sep = "\n")
if(!is.null(NOTE)) cat("\n", "NOTE: ", NOTE, "\n\n", sep = "")
```

y redacta la lista de forma estructurada y clara y centrada en el carácter "=". Algo comparable puede encontrarse en la salida de los cálculos de potencia con `stats:::print.power.htest()`. En términos generales, los criterios de información deben tratarse con precaución, ya que sólo permiten una comparación relativa de los modelos. Los factores de Bayes no hacen otra cosa, que, si están presentes en la escala logarítmica, dan como resultado las diferencias de los criterios de información de los distintos modelos y, en caso contrario, la relación de los modelos. Promediar la Likelihood sobre la Prior (es decir, integrarla sobre los parámetros del modelo sobre los que recaen los supuestos a priori en cada caso) proporciona una protección básica contra el sobreajuste en forma de regularización mínima. Sin embargo, esto no es exactamente lo mismo que la limitación (= penalización) de los valores en *AIC* o *BIC* debido al exceso de parámetros. En el trasfondo, no inesperadamente, la influencia de la Prior a lo largo de todo el proceso de los criterios de información bayesianos, especialmente cuando se trata de la comparación de modelos. El es independiente de si una Prior es informada o no, o se elige por razones de contenido o matemáticas. Por defecto, los criterios de información bayesianos discutidos raramente están equipados con Priores de contenido.

Sin embargo, en lugar de ver los argumentos como un exclusivo PRO vs. CON de los criterios de información clásicos vs. bayesianos McElreath (2015, p.192) subraya,

> „It's important to realize, though, that the choice of Bayesian or not does not also decide between information criteria or Bayes factors. Moreover, there's no need to choose, really. We can always use both and learn from the ways they agree or disagree."

Resumimos: básicamente, todos los criterios de información tratan de resolver la contradicción entre sobreajuste e infraajuste.

6.8.3 Sobreajuste y subajuste

La sobreadaptación y la subadaptación (McElreath, 2015, cap. 6) se refieren a los extremos del rendimiento del modelo cuando la tarea consiste en ajustar un modelo a los datos existentes y probar el poder predictivo del modelo estimado para los nuevos datos.

- El **sobreajuste** se refiere al hecho de que un modelo está sobreajustado a los datos reales y presumiblemente tiene poco poder explicativo nuevo ante los nuevos datos. Técnicamente, esto significa que se han incluido demasiados parámetros no explicativos en el modelo. Estos parámetros deben eliminarse y los criterios de información que tienen esto en cuenta penalizan estos parámetros adicionales no explicativos.
- La **subajuste** se refiere a lo contrario, es decir, que un modelo sólo puede explicar inadecuadamente un conjunto de datos concreto porque faltan las variables explicativas pertinentes. La solución óptima a este problema se reserva, como tantas otras veces, al caso concreto y consiste en encontrar la combinación de variables que realmente explican el modelo.

Ambas situaciones problemáticas reflejan principalmente no sólo problemas estadísticos, sino también consideraciones fundamentales sobre el modelo y la forma de incorporar parámetros al mismo. Por supuesto, las matemáticas desempeñan un papel fundamental en la identificación e inclusión de nuevos parámetros. Pero la matemática no puede decidir qué parámetros deben incluirse, si posiblemente los datos no están disponibles y/o su contenido sólo se conoce de forma limitada. En la entrada del blog (CrossValidated, usuario 3851283, 2014) se discuten ejemplos reales con código R. El siguiente código R demuestra ambas situaciones.

Para el caso de sobreajuste, basta una demostración con predictores polinómicos de orden superior. En primer lugar, se crean los vectores *x* e *y* con números 1:10 y se añade al vector *y* algo de ruido distribuido normalmente. Esto se representa gráficamente. Ahora, sucesivamente para polinomios de orden superior se crean sucesivamente gráficos mediante regresión polinómica. Una regresión polinómica es una regresión en la que un modelo se amplía sucesivamente con predictores de mayor potencia y se obtiene la predicción de la variable dependiente como suma de estos predictores exponenciados. En la versión clásica, esto tiene el siguiente aspecto:

$$y = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + \beta_3 x^3 + ... + \beta_n x^n + \epsilon \qquad (6.74)$$

Obviamente, los polinomios se ajustan cada vez mejor a los datos a medida que aumenta el número (potencias crecientes). Sin embargo, como en este caso el modelo correcto es una simple línea recta, se produce un sobreajuste del modelo a los datos. En el caso de aplicar este modelo a nuevos datos, los datos, probablemente se obtendría un ajuste mucho peor que el que permitiría un modelo más simple. A la inversa, el modelo más simple se comportaría algo peor que el modelo sobreajustado para los datos en cuestión, pero en cambio tendría un poder predictivo significativamente mayor (`ptII_quan_Bayes_over-and-underfitting.r`, véase la Fig. 6.34).

```
# overfitting with polynomials
set.seed(2836)
x <- 1:10
y <- 1:10
y <- y + rnorm(10, 1, 2)
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l")
# plot polynomials
par(oma=c(2,1,4,1), "cex.axis"=1, bty="l", mfrow=c(3,3))
plot(x,y, col="violetred3", pre.plot=grid(), pch=21, cex=1.5,
       bg="darkred", bty="n", main="Scatterplot")
for(i in 1:8)
{
plot(x,y, col="violetred3", pre.plot=grid(), pch=21, cex=1.5,
       bg="darkred", bty="n", main=paste("poly degree = ",i,sep=""))
lines(x,predict(fit <- lm(y ~ poly(x,i))), col="blue")
```

```
summary(fit)
}
mtext("Polynomial (over-)fitting", outer=TRUE, line=0.7, cex=1.5, side=3)
```

Para el subajuste, se toman los números 1 : 100 para los vectores *x* e *y*. Ahora se añade al vector *y* un poco de ruido distribuido normalmente y luego se logaritma *y*, es decir, se cambia la escala. Obviamente, un simple gráfico de un modelo lineal no encaja aquí, como puede verse. Permanezcamos en el plano del modelo lineal, la transformación de *x* a `log(x)` llevaría de nuevo a un ajuste perfecto, al igual que la transformación de *y* (= en la escala `log()`) mediante `exp(y)`.

```
# seed
seed <- 2836
# example 1 - underfitting
set.seed(seed)
x <- 1:100
y <- log(x)
# y <- log(1:100 + rnorm(10, 4, 6))
# create models
fit0.linear <- lm(y ~ x)
fit0.exp <- lm(exp(y) ~ x)
fit0.log <- lm(y ~ log(x))
# plot
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l")
plot(x,y, col="darkred", pre.plot=grid(), pch=21, cex=1.3,  bg="yellow", bty="n")
lines(x,y, col="violetred3", lwd=1, lty=2)
lines(x, predict(fit0.linear), col="yellowgreen", lwd=2)
mtext("Underfitting", outer=TRUE, line=-1.5, cex=1.5, side=3)
par(fig=c(0,1,0,1), oma=c(1,0,0,0), mar=c(0,0,0,0), new=TRUE)
plot(1, type="n", bty="n", xaxt="n", yaxt="n")
legend("bottom", legend=c(expression(paste(log(x))),"linear"),
       lty=c(2,1), lwd=c(1,2), xpd=TRUE, horiz=TRUE,
       col=c("violetred3","yellowgreen","steelblue"), bty="n", cex=.9)
```

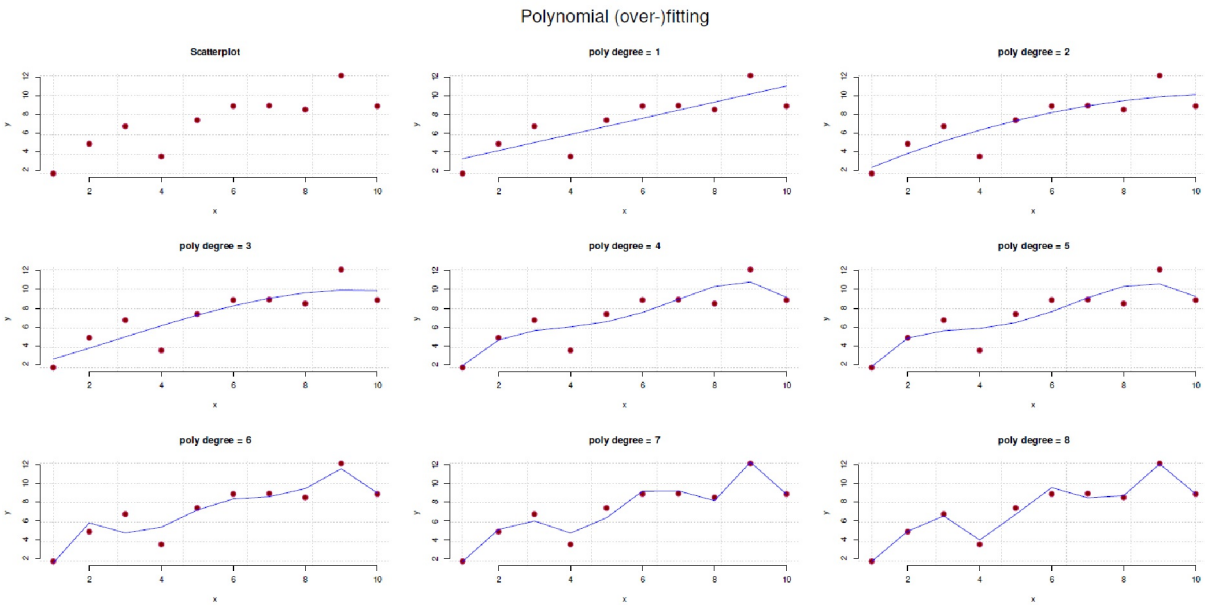


Figura 6.34. *Sobreajuste con regresiones polinómicas*

También se podría generar el caso de forma que el vector *y* se genere en lugar de por un `log(y)` por la suma de polinomios de orden superior de *x*. Ambos casos se imprimen en la Figura 6.35.

```
# example 2 - underfitting R-Code
set.seed(seed)
x <- sort(rnorm(100, 1, 2))
y <- x + x^2 - x^3 + x^4
# create models
fit.linear <- lm(y ~ x)
fit.poly <- lm(y ~ poly(x,3))
# plot
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l")
plot(x,y, col="darkred", pre.plot=grid(), pch=21, cex=1.3,  bg="yellow", bty="n")
lines(x,y, col="violetred3", lwd=1, lty=2)
lines(x, predict(fit.linear), col="yellowgreen", lwd=2)
lines(x, predict(fit.poly), col="steelblue", lwd=2, lty=1)
mtext("Underfitting and polynomial fit", outer=TRUE,
       line=-1.5, cex=1.5, side=3)
par(fig=c(0,1,0,1), oma=c(1,0,0,0), mar=c(0,0,0,0), new=TRUE)
plot(1, type="n", bty="n", xaxt="n", yaxt="n")
legend("bottom",
       legend=c(expression(paste(x^1+x^2-x^3-x^4)),"linear","poly(x,3)"),
       lty=c(2,1,1), lwd=c(1,2,2), xpd=TRUE, horiz=TRUE,
       col=c("violetred3","yellowgreen","steelblue"), bty="n", cex=.9)
```

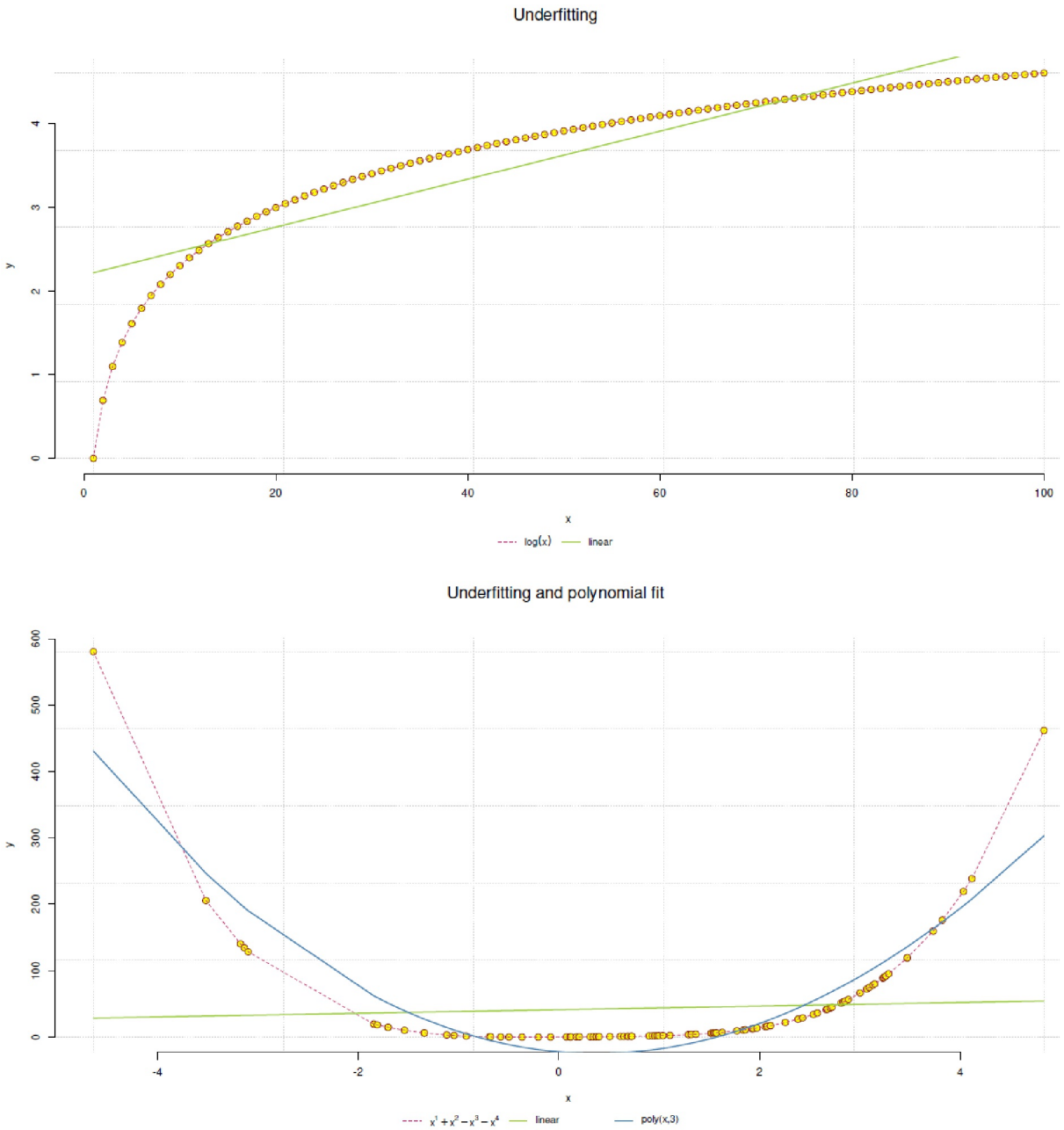


Figura 6.35. *Ajuste insuficiente con regresión simple y polinómica*

Un modelo lineal simple daría un ajuste muy pobre a los datos en ese caso. Cambiando el predictor *x* por polinomios de orden superior con `poly(x,n)` cuando *n* es el grado de orden superior, se obtiene de nuevo un ajuste muy bueno. Sin embargo, es peor en comparación con un `exp(y)` del vector *y*. Se pueden obtener los modelos con `summary()` y luego compararlos. Es impresionante si se toma una simple curva senoidal, que muestra como las estimaciones de los modelos descritos llegan a resultados diferentes (véase Fig. 6.36):

```
# example 3 - underfitting
# sinus curve
set.seed(seed)
x <- sort(rnorm(100, 1, 2))
y.sin <- sin(x)
fit.sin <- lm(y.sin ~ x)
fit.sin.poly5 <- lm(y.sin ~ poly(x,5))
summary(fit.sin)
summary(fit.sin.poly5)
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l")
plot(x,y.sin, col="darkred", pre.plot=grid(), pch=21, cex=1.3, bg="yellow", bty="n")
lines(x,y.sin, col="violetred3", lwd=1, lty=2)
lines(x, predict(fit.sin), col="yellowgreen", lwd=2)
# with poly = 5 it almost works
lines(x, predict(fit.sin.poly5), col="steelblue", lwd=2, lty=1)
mtext("Underfitting and polynomial fit", outer=TRUE, line=-1.5, cex=1.5, side=3)
par(fig=c(0,1,0,1), oma=c(1,0,0,0), mar=c(0,0,0,0), new=TRUE)
plot(1, type="n", bty="n", xaxt="n", yaxt="n")
legend("bottom", legend=c("sinus","linear","poly(x,5)"),
       lty=c(2,1,1), lwd=c(1,2,2), xpd=TRUE, horiz=TRUE,
       col=c("violetred3","yellowgreen","steelblue"), bty="n", cex=.9)
```

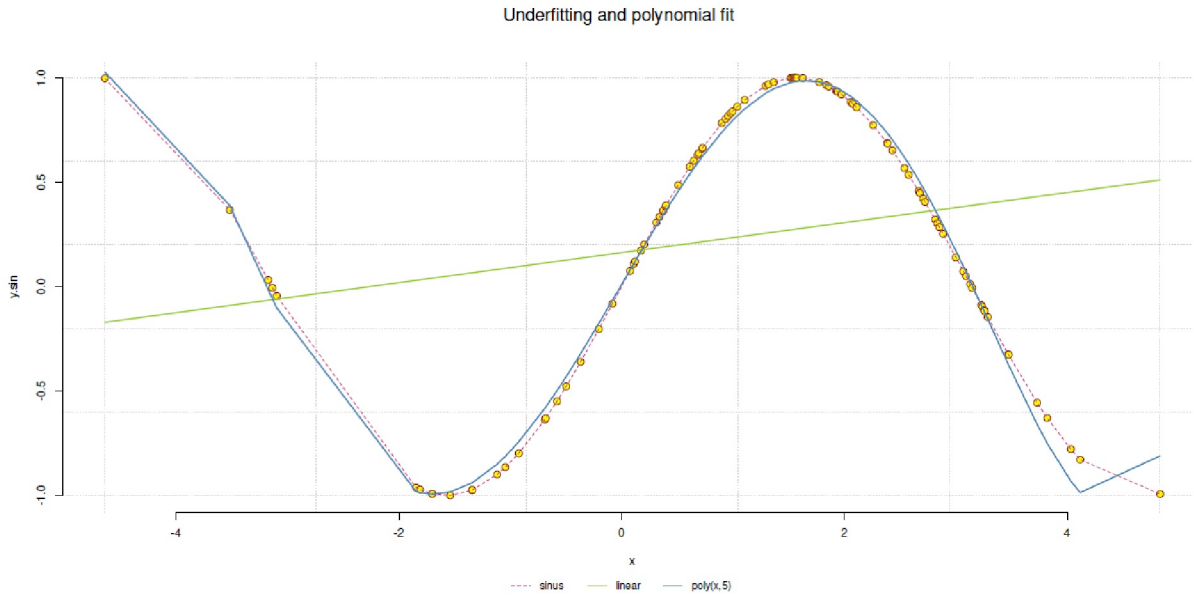


Figura 6.36. *Ajuste insuficiente con regresión simple y polinómica (curva sinusoidal)*

Esto debería demostrar que siempre son sólo modelos y que no existe el mejor modelo definitivo en todas las circunstancias.

Otra forma de entender la diferencia entre ajuste excesivo y ajuste insuficiente es imaginar un gran conjunto de datos, por ejemplo, los datos del Titanic en los capítulos 5.5.4 y 12.11.2. Ahora creamos un modelo en una parte del conjunto de datos, por ejemplo, para predecir la supervivencia. Lo probamos críticamente en la otra parte del conjunto de datos en términos de predicción. Del mismo modo, el modelo podría predecir datos generados por simulación sobre la base del modelo (véase también bootstrap en el capítulo 4.3.5 o comprobaciones predictivas posteriores en el capítulo 6.8.4.3). Simplificando, existe entonces una curva generada por el modelo para los datos de entrenamiento y otra para los datos de prueba. En caso de subajuste, la curva tiene un error elevado tanto para los datos de entrenamiento como para los datos de prueba debido a la estimación del modelo, siempre que ambos sean en principio iguales. En cambio, en el caso de sobreajuste, el conjunto de datos de entrenamiento se estima muy bien y el conjunto de datos de prueba muy mal. Utilizando el ejemplo de los polinomios anterior, queda más claro (o no) por qué

demasiados de ellos conducen al sobreajuste. En este caso, el aprendizaje no se produce para combinar modelo y datos hacia un ajuste óptimo, sino sólo – con un número creciente de predictores polinómicos – un "recuerdo" de los datos. Naturalmente, esto falla con datos nuevos porque no hay nada que recordar y no se ha reconocido la estructura inherente a los datos. La tarea del modelo no es recordar, sino descubrir las estructuras básicas de los datos y sus relaciones mutuas para poder transferirlas a nuevas situaciones (datos). Esto significa detener la adaptación a los datos de entrenamiento en un punto determinado para permitir una flexibilidad básica para los datos de prueba. De lo contrario, el modelo estimado será demasiado rígido y dejará de ser lo suficientemente flexible para la predicción.

El ajuste de modelos desempeña un papel tan importante porque la investigación debería centrarse siempre menos en si los modelos son "*correctos*" o "*incorrectos*" (lógica de "lo uno o lo otro") y más en comprender en qué momento el modelo *no puede* explicar los datos. Ahí es donde se puede aprender y mejorar el modelo. Existen diferentes estrategias para el ajuste de modelos. La mejor es encontrar un modelo sustantivo que se ajuste específicamente al contexto y a la pregunta de investigación. Una variante pragmática pero menos útil sería generar modelos con, por ejemplo, polinomios de orden superior y *por fuerza bruta* hasta que el error en la predicción de los datos de entrenamiento *y* los datos de prueba está minimizado. Esto corresponde a un problema de optimización o a una función de pérdida minimizada (véase el capítulo 6.8.1.2) y pertenece al ámbito de la validación cruzada. Además, puede combinarse con la omisión selectiva de los datos de entrenamiento o de prueba (por ejemplo, validación cruzada leave-one-out). Un ejemplo de validación cruzada lo ofrece Koehrsen (2018) en una entrada de blog. Aunque este enfoque puede conducir a buenos resultados numéricos en muchos casos, sigue habiendo una gran incertidumbre sobre lo que significa el modelo en términos de contenido aparte de la suma de predictores, cada uno con una potencia diferente, que en la mayoría de los casos tiene poco poder explicativo en términos de contenido.

En este sentido, la variante de un modelo *basado en el contenido* debería ser definitivamente perseguida. Por ejemplo, una combinación inteligente de funciones no lineales puede ser más valiosa en términos de contenido y claramente superior a la simple combinación de polinomios (Stansbury, 2013). Por supuesto, siempre es útil disponer de una gran base de datos, especialmente si es mayor que el conjunto de datos de prueba, lo que hace que la generalización sea más factible. Los modelos en sí mismos pueden entenderse (McElreath, 2015, p.172) como compresiones que pueden reducir conjuntos de datos muy grandes con muy pocos parámetros y describirlos con la mayor precisión posible. El paradigma de codificación (véase el capítulo 9) no hace otra cosa a través de la codificación en la investigación cualitativa. Lo mismo se aplica a la hipótesis de la estructura de casos del análisis secuencial de la hermenéutica objetiva (véase el capítulo 11).

La *regularización* es otro aspecto de la comprobación de modelos, con el fin de no utilizar supuestas características del modelo estimado, contrarrestando así el sobreajuste. En este proceso, se poda el modelo y, por ejemplo, no se utiliza cierta información o sólo se utiliza parcialmente, se limita el ancho de banda, etc. A continuación, el modelo se utiliza de la misma forma que el modelo estimado. El motivo es lograr que el modelo se base en la información principal que tiene más posibilidades de generalización. Esto requiere reflexión. El enfoque es similar en idea a lo que se hace en el análisis robusto de datos. Se puede encontrar un ejemplo en R en la viñeta del paquete R `keras` (Falbel et al., 2019) en el contexto del aprendizaje automático. Esto revela inmediatamente el peligro de los análisis automatizados, contra los que advierten Gigerenzer y Marewski (2015): una máquina está lejos de ser capaz de distinguir lo que es significativo o no en términos de contenido. No tener esto en cuenta conduce inevitablemente a conclusiones incoherentes. Especialmente en modelos complejos con muchos parámetros, existe un peligro especial de sobreajuste (por ejemplo, los modelos de "random forest" o las redes neuronales en el aprendizaje automático). La complejidad y la sobreadaptación están estrechamente relacionadas, al igual que la complejidad y la inadaptación: en el caso de la sobreadaptación, debe reducirse la complejidad, mientras que en el caso de la subadaptación, debe aumentarse la complejidad. Caso por caso, hay una zona en el que ambos se equilibran, y es ahí donde suele encontrarse el modelo óptimo.

Epistemológicamente, por un lado, la navaja de Occam está disponible para simplificar de forma limitante los modelos sobreajustados demasiado complejos. Por otro lado, la simplificación excesiva conduce a la falta de aprendizaje y a la ausencia de conocimiento. El término medio equilibra cuidadosamente ambas

partes. En una profundización de este complejo problema, el tema conduce, por un lado, a la maximización de la entropía (= máxima entropía, véase cap. 6.14) para representar adecuadamente determinados estados de información. Por otro lado, conduce a la comparación de distribuciones, por ejemplo mediante las *divergencias de Kullback-Leibler* o *entropía/información* (Kullback & Leibler, 1951). También se encuentra el nombre de *entropía relativa* para estudiar los sistemas de información mediante la entropía de Shannon (1948). La divergencia de Kullback-Leibler compara dos distribuciones de probabilidad, una de las cuales sirve de referencia (por ejemplo, observación frente a modelo o estimación de modelo frente a predicción con datos nuevos o simulados). Para valores discretos, la divergencia KL se calcula con las distribuciones discretas *P* y *Q* a

$$D_{\mathrm{KL}}(P \,||\, Q) = \sum_{x \epsilon X}^{\infty} P(x) \cdot log \left[\frac{P(x)}{Q(x)} \right] \tag{6.75}$$

y para valores continuos con las funciones de densidad de probabilidad `p(x)` y `q()`

$$D_{\mathrm{KL}}(P \,||\, Q) = \int_{-\infty}^{\infty} p(x) \cdot log \left[\frac{p(x)}{q(x)} \right] \mathrm{d}x \tag{6.76}$$

En R, la divergencia de Kullback-Leibler se encuentra en los paquetes de R `entropy` con `KL.plugin()`, en `catR` con `KL()` para modelos IRT y varias funciones R en `entropyMCMC`. El ejemplo de `KL.plugin()` facilita la comprensión de su funcionamiento (`ptII_quan_Bayes_over-and-underfitting.r`):

```
seed <- 83756
set.seed(seed)
# create some counts
freq1 <- runif(10,0,11)
freq2 <- runif(10,0,20)
freq3 <- runif(20,0,20)
freq4 <- runif(20,0,10)
freq5 <- runif(10,0,11)
freqlist <- list(freq1,freq2,freq3,freq4,freq5)
freqlist
# create probabilities from frequencies
flist <- lapply(freqlist, function(x) x/sum(x))
flist
# sanity check whether sum up to 1
lapply(flist, sum)
chooses <- combn(length(flist),2)
chooses
rownames(chooses) <- c("freq#1","freq#2")
# calculate Kullback-Leibler divergence between the probabilities
KL.plugin.res <- apply(chooses,2,function(x)
KL.plugin(freqlist[[x[1]]], freqlist[[x[2]]]))
```

La función de R `combn()` ayuda enormemente a comprobar todos los elementos de la lista generada entre sí. Ahora, como comprobación cruzada, escribimos nuestra propia función `KL.man()` para determinar manualmente la divergencia de Kullback-Leibler según la fórmula anterior para valores discretos

```
# manual calculation
KL.man <- function(pe,qu)
{
pe <- pe/sum(pe)
qu <- qu/sum(qu)
# if p((x)) <= 0 -> corresponding term equals zero,
# because lim x*log(x) = 0 for x<0
logfac <- ifelse(pe>0, log(pe/qu), 0)
```

```
KL <- sum(pe * logfac)
return(KL)
}
```

aplicamos la función

```
# apply it R-Code
KL.man.res <- apply(chooses,2,function(x)
KL.man(freqlist[[x[1]]], freqlist[[x[2]]]))
```

y comparamos los resultados:

```
> # compare
> all.equal(KL.plugin.res,KL.man.res)
[1] TRUE
```

Son iguales, como era de esperar. Además, se pueden obtener χ^2-estadísticas:

```
> # calculate chi^2-statistic between the probabilities
> chi2 <- apply(chooses,2,
function(x) chi2.plugin(freqlist[[x[1]]], freqlist[[x[2]]]))
> chi2half <- chi2/2
> comparisontab <- t(rbind(chooses,KL.plugin.res,
KL.man.res,chi2,chi2half))
> round(comparisontab,3)
       freq#1 freq#2 KL.plugin.res KL.man.res chi2   chi2half
[1,]    1      2       0.696        0.696      3.353  1.676
[2,]    1      3       1.076        1.076     12.688  6.344
[3,]    1      4       0.900        0.900      3.831  1.915
[4,]    1      5       0.394        0.394      2.393  1.197
[5,]    2      3       0.947        0.947      5.818  2.909
[6,]    2      4       1.080        1.080      4.429  2.214
[7,]    2      5       0.462        0.462      1.408  0.704
[8,]    3      4       0.411        0.411      1.369  0.685
[9,]    3      5      -0.194       -0.194      1.460  0.730
[10,]   4      5      -0.280       -0.280      1.748  0.874
```

La situación descrita hasta ahora para las pruebas de hipótesis bayesianas lleva a centrarse cada vez más en las alternativas. Esto incluye el *ajuste de modelos*, la *comprobación de la realidad* de los modelos mediante *simulación* y *replicación*, y transversalmente, ya que es constantemente necesario, la *evaluación gráfica* de los modelos. No se trata de aceptar o rechazar hipótesis y modelos, sino de *comprender* qué ocurre cuando tal o cual modelo se combina con los datos en tal o cual variante.

6.8.4 Modelización y estimación por intervalos

6.8.4.1 Estimación bayesiana por intervalos

Cada estimación de modelo bayesiano da como resultado una estimación de los parámetros desconocidos con una distribución de probabilidad posterior asociada. Ésta puede representarse gráficamente y muestra la distribución del parámetro en función de sus valores. A partir de ella se pueden derivar intervalos de confianza bayesianos (= *intervalos creíbles*, abreviados aquí como CI_{Bayes}), que indican con qué (in)certeza se encuentra un parámetro en un rango determinado. De este modo, se puede calcular la probabilidad de

un determinado intervalo de valores o bien se selecciona el grado de certeza y luego se examina para ver lo grande o pequeño que parece el intervalo de valores resultante y qué valores se encuentran aquí en el rango superior e inferior.

Estos intervalos de confianza difieren de los intervalos de Neyman-Pearson (véase el capítulo 4.3.3), ya que indican en realidad la probabilidad con la que un parámetro se encuentra en un intervalo de valores. En cambio, los intervalos de Neyman-Pearson se basan en mediciones de muestras aleatorias repetidas e indican la probabilidad de que los datos se ajusten a la hipótesis investigada. Técnicamente, el intervalo de confianza de Neyman-Pearson dice algo sobre el intervalo en el que se situarán los valores con confianza (1-α)% y con mediciones repetidas. El intervalo no puede predecir nada sobre mediciones concretas, ya que no hace ninguna afirmación sobre la probabilidad con la que un parámetro se encuentra en un intervalo de valores que se debe definir. Los valores concretos siempre se encuentran o no en el intervalo de confianza, entre ambos no hay zona gris.

La salida gráfica de la distribución posterior también es importante para investigar si la distribución posterior es simétrica en torno al máximo o sesgada. Esto tiene diferentes consecuencias, en primer lugar con respecto a la cuestión del ajuste adecuado del modelo y luego con respecto a la estimación bayesiana de intervalos. Mientras que un CI_{Bayes} asume colas igualmente grandes al final de la distribución posterior, esto no se aplica a los *high density intervals* (= HDI; intervalos de alta densidad), para los que las colas no tienen que ser igualmente grandes. Un HDI es el CI_{Bayes} más corto a partir del valor modal de la distribución. Es característico que todos los valores dentro del HDI sean mayores que los valores fuera del HDI, es decir, los valores en el HDI tienen las mayores probabilidades. La masa del HDI corresponde a el (1-α)% elegido, por ejemplo "típicamente" 95% – donde estamos de nuevo con convenciones. Otra ventaja sobre los CI_{Bayes} es que los HDI pueden utilizarse tanto para distribuciones simétricas como no simétricas (por ejemplo, sesgadas) unimodales e incluso bimodales (Kruschke, 2015b, p.88). En el caso de una distribución bimodal, el HDI se divide en dos subintervalos, uno para cada uno de los dos valores modales. El cálculo se realiza entonces de forma equivalente al caso unimodal y se aplican las mismas condiciones (los valores dentro del HDI son mayores que los valores fuera y la masa total del HDI corresponde al x%). Los HDI también se conocen como *high density posterior intervals* (= HPD; intervalos de alta densidad posterior). La distinción estricta entre los términos "HDI" y "HPD" sólo indica que los HDI abarcan también las Priors, etc., mientras que los HPD, en sentido estricto, sólo se refieren a las Posteriors. El cálculo es idéntico. La amplitud o estrechez de los HDI proporcionan información sobre la certeza de las estimaciones. Un intervalo amplio que cubre un gran espectro del parámetro θ de interés tiene una incertidumbre mayor que un intervalo estrecho que se limita a una pequeña parte del espectro. Los HDI pueden calcularse en R utilizando `hdi()` del paquete `HDInterval` de R para una variedad de objetos como vectores, densidades, tablas o funciones u objetos MCMC. `hdi()` es fácil de usar siempre y cuando uno sea consciente de que no es un intervalo simétrico (`ptII_quan_Bayes_HDI.r`):

```
# example how to use hdi()
> prob <- 0.87
> seed <- 0987
> set.seed(seed)
> rn <- rnorm(100,10,2.5)
> hdi(rn, credMass=prob)
lower      upper
6.755363   14.259990
attr(,"credMass")
[1] 0.87
```

Un ejemplo de la diferencia entre CI_{Bayes} y el HDI lo demuestra Kruschke (2012b). Ficticiamente, el lanzamiento de una moneda dio *diez caras en diez ensayos*. Si esto se modela con una probabilidad de Bernoulli y una distribución uniforme como Prior, es decir, `dbeta(1,1)`, el resultado es una Posterior de `dbeta(1,11)` (véase la Fig. 6.37).

```
# example from Kruschke (2012) R-Code
prob <- 0.95
```

```
theta <- seq(0,1,0.001)
randist <- dbeta(theta, shape1=1, shape2=11)
dens <- list(x=theta, y=randist)
attr(dens, "class") <- "density"
```

Como CI_{Bayes} y HDI resultan

```
> plotHDI(dens=dens, prob=prob, quants=qbeta(c(0.025,0.975),1,11),
densTF=TRUE, digs=4)
############################
Interval    [0.025; 0.975]
HDI         [0; 0.238]
CI (sym)    [0.0023; 0.2849]
############################
```

Obviamente, el valor modal es cero. Un CI_{Bayes} típico del 95%, dado que corta estoicamente el 2.5% de ambos extremos, cortaría el valor modal y, por tanto, excluiría el valor cero. En cambio, un intervalo de HDI al 95% no hace esto e incluye aquí el valor modal cero. Los intervalos de confianza resultantes difieren. El CI_{Bayes} simétrico al 95% excluye el cero y va de 0.0023 a 0.2849, mientras que el HDI al 95% va de 0 a 0.238. El HDI al 95% sólo excluiría el cero si la propia Prior excluyera el cero como valor, ya que (véase el capítulo 6.12 sobre regularización) la Posterior es el producto de la Prior * Likelihood; y si un factor de un producto es cero, [...] etc. Por otra parte, una desventaja de los HDI es que no son invariantes frente a transformaciones no lineales de las variables (Kruschke, 2015, p.342f.). Los CI_{Bayes} permanecen invariantes. Esto significa que una transformación, por ejemplo un escalado, tiene un efecto negativo y distorsiona los límites de los HDI. Sin embargo, los intervalos suelen interpretarse en determinadas escalas para que sean significativos en términos de contenido, pueden surgir problemas de interpretación. Kruschke (2015, p.343) señala,

> „This property [= HDI, Nota de los autores] is handy when the parameters are arbitrarily scaled in abstract model derivations, or in some applied models for which parameters might be nonlinearly transformed for di erent purposes. But in most applications, the parameters are meaningfully defined on the canonical scale of the data, and the HDI has meaning relative to that scale. Nevertheless, it is important to recognize that if the scale of the parameter is nonlinearly transformed, the HDI limits will change relative to the percentiles of the distribution."

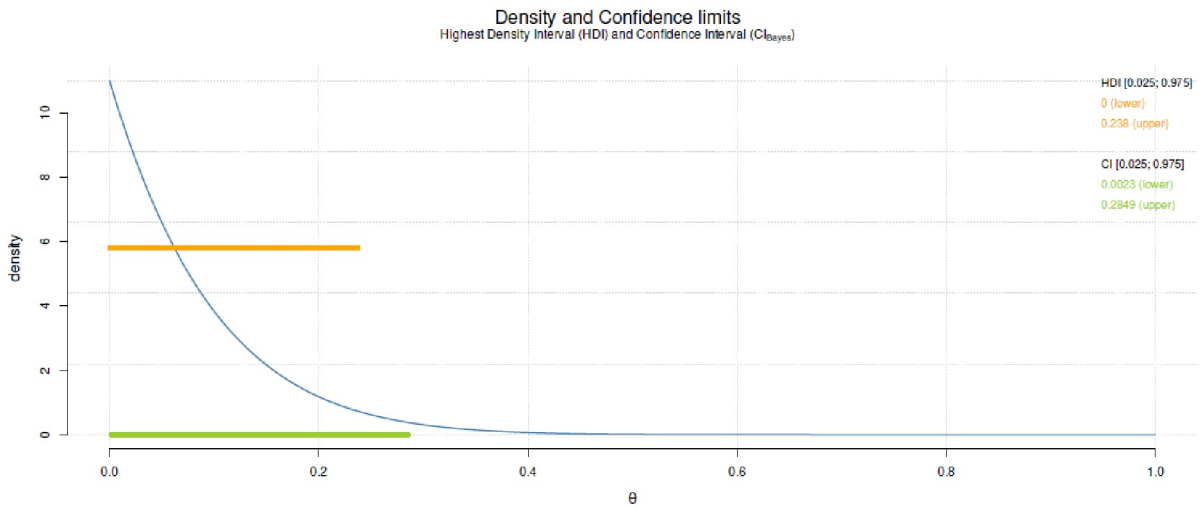


Figura 6.37. *HDI vs. CI_{Bayes} (Prior uniforme)*

El siguiente código R con la función R `plotHDI()` demuestra la diferencia entre un HDI y un CI_{Bayes} a lo largo de diferentes valores iniciales (valores aleatorios y densidades de la distribución normal o beta, véase la Fig. 6.38).

```
# plot various HDIs and CIs (Bayes)
# prob of intervall
prob <- 0.87
set.seed(seed)
# create some data / density
theta <- seq(0,1,0.001)
randist <- rnorm(1e5, 2, 1.23)
plotHDI(dens=randist, prob=prob, densTF=FALSE)
# create some data / density
randist <- rbeta(1e5, shape1=2.5, shape2=8)
plotHDI(dens=randist, prob=prob, densTF=FALSE)
```

Por lo general, los valores pueden generarse a partir de la distribución posterior mediante simulación MCMC, como se verá más adelante con las comprobaciones posteriores predictivas (véase cap. 6.8.4.3). A continuación, se trata de la evaluación de la calidad del modelo.

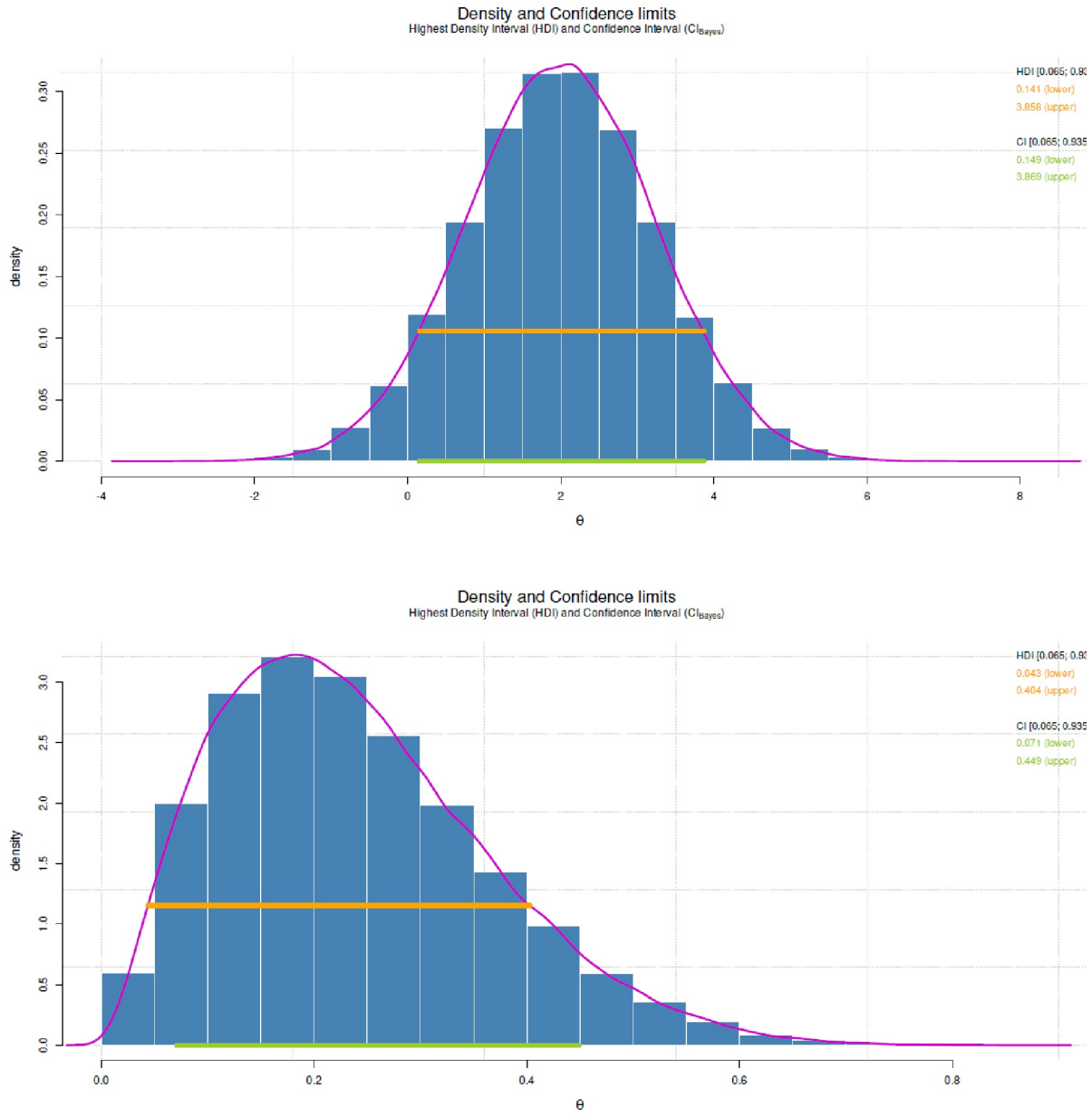


Figura 6.38. *HDI vs.* CI_{Bayes}

6.8.4.2 ROPE – Tolerancia en la estimación puntual

En el contexto de la estimación bayesiana por intervalos y las pruebas de hipótesis, Kruschke (2015) sugiere denotar una región de equivalencia práctica para probar hipótesis nulas puntuales de modo bayesiano. Una hipótesis nula puede entonces ser probable o incluso bastante inequívoca, si una HDI del 89% (¡o un valor distinto del 89%!) de la distribución posterior del parámetro de interés cubre el valor cero. Pero también es posible que el valor cero no esté enmascarado en absoluto. En tercer lugar, existe una zona gris en la que

podría ser, con cierto grado de incertidumbre que el cero esté superpuesto, sólo que no lo sabemos con exactitud. Mientras que las dos primeras situaciones permiten tomar una decisión clara, esto ya no es posible en la zona gris. Se impide tomar una decisión clara, lo que sin duda ocurre más a menudo en la práctica. En consecuencia, se debe revisar el diseño, los datos, el modelo, etc. para averiguar cuáles son las razones de esta situación ambigua.

Sin embargo, para manejar mejor las tres situaciones desde un punto de vista técnico y evitar caer en los abismos de la estadística clásica y el manejo de las barreras de significancia, Kruschke (2015) propone definir una *región crítica* alrededor del valor a probar – en este caso el valor cero, pero puede ser cualquier parámetro θ – que se identifica a causa de efectos prácticos, es decir, se le equipara como el valor cero o el parámetro θ. Por tanto, se asigna el valor cero a un punto, un intervalo de tolerancia, una zona gris, y con este intervalo se realizan las comparaciones anteriores para el solapamiento o no con el x% HDI del parámetro de interés. Se trata de una comparación entre dos intervalos y no entre estimaciones por intervalos y por puntos.

Esta extensión del valor puntual (valor cero, parámetro, ...) es la región de equivalencia práctica o ROPE (= Region Of Practical Equivalence). Kruschke (2015, p.336, cursiva en el original) la designa de la siguiente manera:

> „A region of practical equivalence (ROPE) indicates a small range of parameter values that are considered to be practically equivalent to the null value for the purposes of the practicular application."

Esto tiene la ventaja de que no se trata de si, por ejemplo, en la evaluación de un lanzamiento de moneda justo la probabilidad de "cara o cruz" debe ser exactamente p = 0.50. La casi nunca ocurre en la práctica, habida cuenta del tamaño limitado de las muestras y los múltiples errores de medición. Más bien, basta con saber que el valor posterior se aproxima lo suficiente a p = 0.50 para suponer la equidad de una moneda. Es irrelevante que ahora sea $p = 0.483$, $p = 0.523$, etc. Por lo tanto, se indica un intervalo, por ejemplo $0.45 < p < 0.55$ para muestras pequeñas o $0.49 < p < 0.51$ para muestras mucho más grandes, que describe el ROPE. A la inversa, pedimos efectos en lugar de hipótesis nulas, el ROPE es útil para decir algo sobre si un efecto está al menos, por ejemplo, un 7% puntos por encima o por debajo (dependiendo de la dirección de la hipótesis) de la hipótesis nula (= ningún efecto), es decir, ROPE = ∓7% puntos en relación con el parámetro θ. La idea subyacente es que si el x% HDI de un parámetro está completamente fuera del ROPE se ha dejado la zona gris y es claramente más probable que se produzca un efecto, una diferencia de cero, etc. El ROPE no resuelve realmente el problema de determinar los límites, sino que subdivide el área crítica de una manera que hace que la decisión a favor o en contra de un efecto sea mucho más fácil que un límite puntual como $p = 0.05$ en estadística clásica, donde $p = 0.49$ se considera significativo, mientras que p = 0.51 se declara no significativo. Hay un título muy apropiado de un artículo de revista de Gelman y Stern (2006), que reza "The Difference Between 'Significant' and 'Not Significant' is not Itself Statistically Significant". Gelman (2015d, 2015b) ofrece un un ejemplo práctico de esto.

Utilizando el ejemplo de unos datos distribuidos normalmente con μ = 10 y σ = 2.5, examina con `BESTmcmc()` del paquete R `BEST`, si en la variante de la prueba *t* bayesiana (prueba de una muestra) la media es compatible con un determinado valor (aquí: ficticio) `compVal` = 9. Los valores descriptivos tienen este aspecto, incluyendo la *d* de Cohen y la diferencia de medias en la escala original, es decir, no estandarizada (`ptII_quan_Bayes_ROPE-BayesFactor.r`).

```
> # difference in means example 2 - one sample test
> # compared to zero -> effect size d = (10-0)/2.5 = 4
> seed <- 2745
> set.seed(seed)
> n <- 100
> prob <- 0.87
> mu <- 10
> sigma <- 2.5
> norms <- rnorm(n=n, mean=mu, sd=sigma)
```

```
> c(summary(norms),SD=sd(norms),VAR=var(norms))
Min.      1st Qu.  Median    Mean     3rd Qu.    Max.
1.757645 8.149221 10.304962 9.919335 11.884015 16.184574
SD        VAR
2.789831 7.783157
> # comparecrit
> comparecrit <- 9
> # empirical effect size against comparison value (scalar)
> cohensd(s1=norms,s2=comparecrit)
d|one sample s2 [scalar]
0.3295307 9
> # theory
> mu - comparecrit
[1] 1
> (mu - comparecrit)/sigma
[1] 0.4
> hdi(norms, credMass=prob)
lower     upper
6.760185 14.248162
attr(,"credMass")
[1] 0.87
```

Sigue la llamada de `BESTmcmc()` y la salida.

```
mcmc1 <- BESTmcmc(norms) R-Code
summary(mcmc1)
plotAll(mcmc1)
pairs(mcmc1)
```

Hay varias alternativas gráficas (véase la Fig. 6.39), por lo que `BEST:::plotAll()` produce aquí las más importantes – incluyendo densidades de la distribución posterior y la distribución predictiva posterior (véase también el capítulo 6.8.4.3).

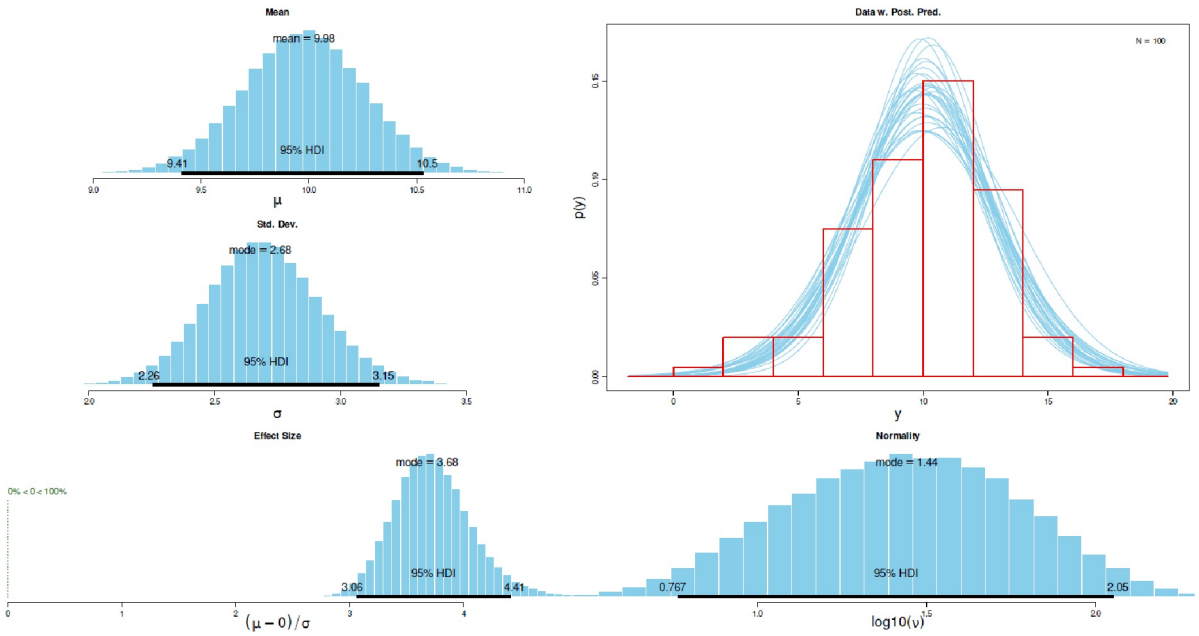


Figura 6.39. BESTmcmc() *(Posterior)*

A continuación, se determina la diferencia con el valor de referencia de interés y qué parte de la masa de la diferencia posterior es mayor que cero. Obviamente, hay un descenso entre 1. y 2. La tarea para el lector interesado consistiría en trazar esta diferencia y averiguar dónde se encuentra exactamente este descenso.

```
> # prob of mu_diff for various values
> mean(mcmc1$mu)
[1] 9.976192
> mudiff <- mcmc1$mu - comparecrit
> mean(mudiff)
[1] 0.9761917
> mean(mudiff > 0)
[1] 0.99958
> mean(mudiff > 0.5)
[1] 0.952231
> mean(mudiff > 1)
[1] 0.4687006
> mean(mudiff > 2)
[1] 0.0001899962
```

Ahora examinamos en el gráfico el ROPE y su alrededor más detalladamente (véase fig. 6.40):

```
sek <- seq(0,2.5,0.01)
probs <- vector()
for(i in 1:length(sek)) probs[i] <- mean(mudiff > sek[i])
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l")
plot(sek,probs, bty="n", col="darkred", type="l", pre.plot=grid(),
       ylab="p", xlab=expression(paste(mu[diff])))
mtext(expression(paste("Probability of different ",mu[diff])),
       outer=TRUE, line=-2, cex=1.5, side=3)
```

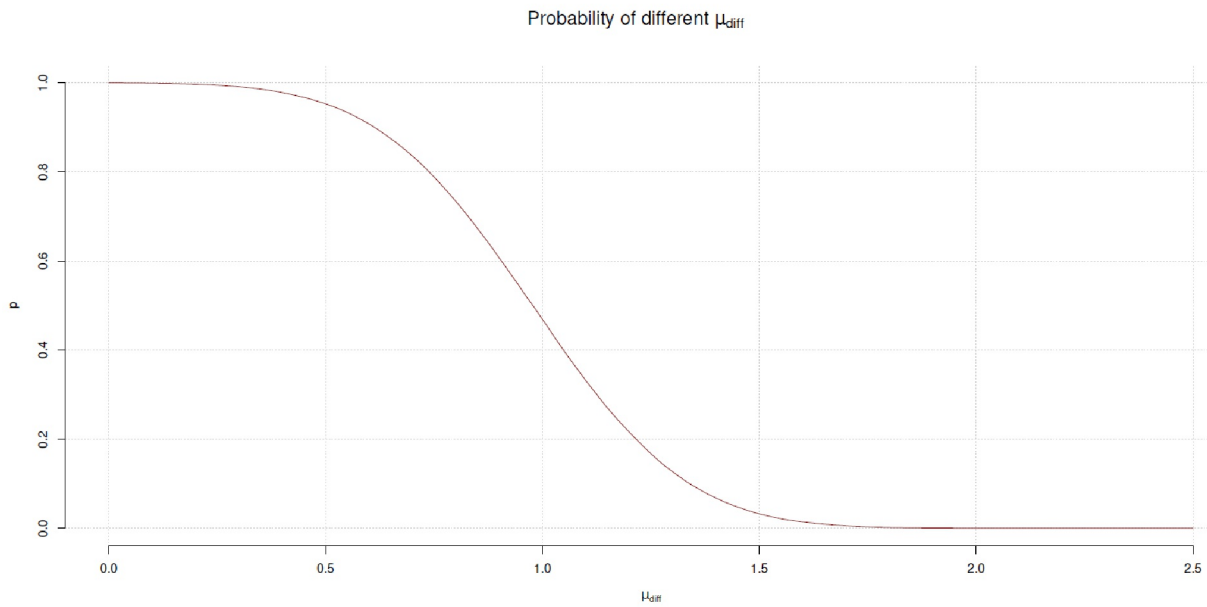


Figura 6.40. BESTmcmc() *(Diferenca del valor de referencia)*

Con `plotAreaInROPE()` se examina qué parte de la masa acumulada de la Posterior se encuentra dentro o fuera de la ROPE en función de la anchura de la ROPE. Si la masa de las Posteriores está completamente dentro de la ROPE, obviamente se ve diferente que si la ROPE se encuentra muy lejos de las Posteriores (para

ejemplos Kruschke, 2013b). Si se trazan los datos originales con `plotAreaInROPE()`, el radio del ROPE debe seleccionarse en esta escala. Si se toman las diferencias μ_{Dif}, la escala de las diferencias se aplica en consecuencia. Si el radio es demasiado pequeño, debe ampliarse. Esto se puede ver en el eje Y cuando la probabilidad allí no llega casi a uno y si el HDI no se muestra en total (véase la Fig. 6.41).

```
# plot ROPE and area-in-ROPE
par(oma=c(2,1,3,1), "cex.axis"=1, bty="l", mfrow=c(2,2))
# mcmc difference in means (against zero!, see above)
plotAreaInROPE(mudiff, credMass=prob, compVal=1.2,
       maxROPEradius=1.5,
       main=expression(paste(mu[diff])))
# raw data
plotAreaInROPE(norms, credMass=prob, compVal=9,
maxROPEradius=8, main="raw")
# posterior
plotPost(mudiff, credMass=prob, compVal=1.2, ROPE=c(0.8,1.4),
       showMode=TRUE, col="grey90", border="white",
       xlab=expression(theta[2]-theta[1]),
       ylab="Density", main=expression(paste(mu[diff])))
lines(density(mudiff), col="violetred3", lwd=2, lty=2)
# posterior + histogram
plotPostPred(mcmc1)
mtext("Area in ROPE", outer=TRUE, line=0, cex=1.5, side=3)
```

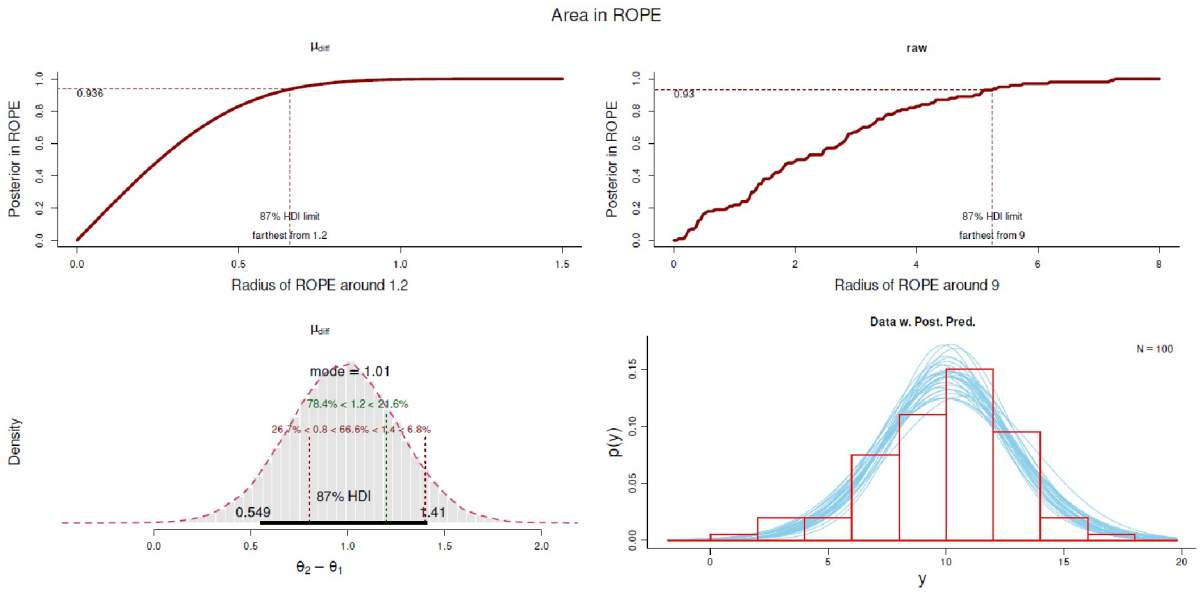


Figura 6.41. *BESTmcmc() (ROPE)*

Esto se puede hacer también para la divergencia media (prueba de dos muestras) o modelos potencialmente más complejos. El establecimiento y cálculo de una ROPE y sus límites ("¿En qué punto, a qué nivel, hablamos realmente de un efecto?"), para determinar el principio y el final o el intervalo permitido del parámetro θ, por ejemplo ROPE = ±5%. No existe una respuesta definitativamente correcta a esta pregunta, mientras que ROPE no se refiere a la exactitud puntual, sino a la exactitud coherente y justificable, que se basa en el objeto y su escala original. Esto permite realizar análisis claramente más exactos. En particular, es posible cambiar el radio de ROPE y trazar los solapamientos respectivos (en puntos porcentuales) con el HDI *x%* del parámetro de interés entre sí. Entonces se obtiene una impresión de los

solapamientos en función del tamaño de la ROPE. Kruschke (2013b) da un ejemplo en su blog mostrando casos en los que, por un lado, la comparación de HDI y ROPE como prueba de parámetros rechaza o no un parámetro, y cómo esto depende de la certeza denotada (amplitud del intervalo) y del tamaño de la ROPE.

A su vez, `plotAreaInROPE()` ofrece exactamente la mayor ROPE posible o su radio alrededor del parámetro examinado marcando el límite del HDI. que todavía puede rechazar el parámetro (ver Fig. 6.42).

```
par(oma=c(2,1,3,1), "cex.axis"=1, bty="l", mfrow=c(2,2))
# area too small
plotAreaInROPE(mudiff, credMass=prob, compVal=1.2,
       maxROPEradius=0.1, main=expression(paste(mu[diff])))
# extend area
plotAreaInROPE(mudiff, credMass=prob, compVal=1.2,
       maxROPEradius=0.5, main=expression(paste(mu[diff])))
# extend it more
plotAreaInROPE(mudiff, credMass=prob, compVal=1.2,
       maxROPEradius=0.8, main=expression(paste(mu[diff])))
# extend it more
plotAreaInROPE(mudiff, credMass=prob, compVal=1.2,
       maxROPEradius=1, main=expression(paste(mu[diff])))
mtext("Area in ROPE - varying radius", outer=TRUE, line=0, cex=1.5, side=3)
```

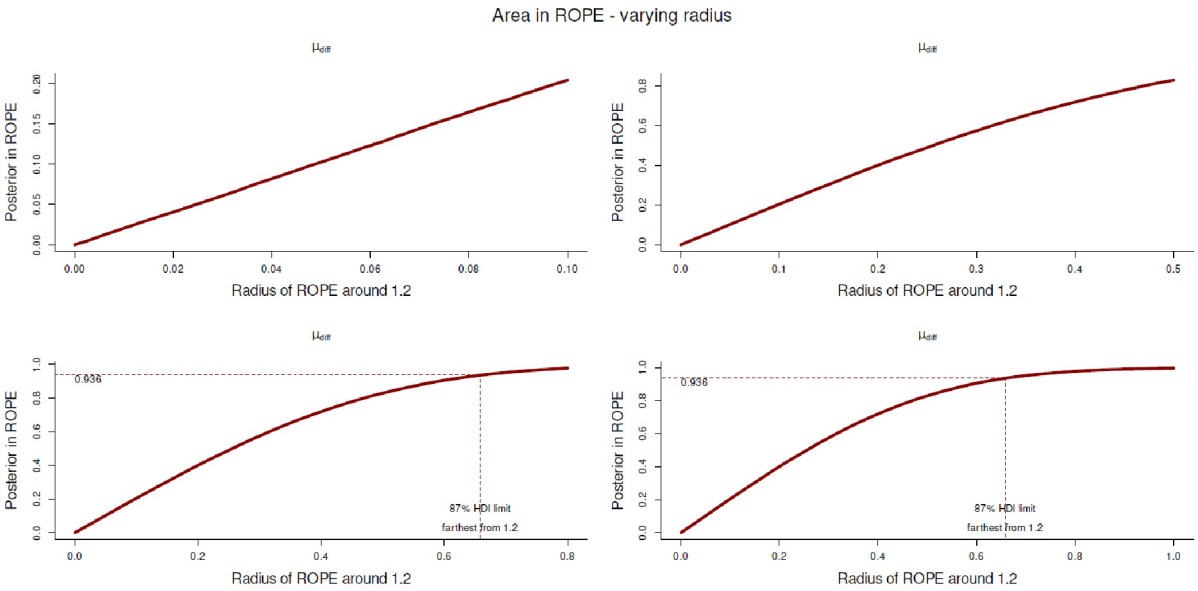


Figura 6.42. *BESTmcmc() (ROPE y radio)*

ROPE, x% HDI y la distribución posterior, así como el parámetro de interés (cero u otro valor) pueden marcarse en un gráfico. Así pues, ROPE es principalmente un procedimiento gráfico. Por supuesto, puede ser útil una consulta lógica o numérica posterior sobre el solapamiento, si el valor entra o no en el ROPE o en la zona gris. Si a partir de ahí se formula una prueba de hipótesis, se produce una aceptación o rechazo de la hipótesis nula de tal forma que, según Kruschke (2015b, p.336), "Por tanto, declaramos que el valor nulo de 0,5 se rechaza a efectos prácticos". La frase "a efectos prácticos" hace hincapié en el carácter práctico y pragmático de este tipo de pruebas, más que en la idea de una prueba universalmente válida. También es importante señalar que se trata de una prueba dirigida al parámetro θ *específico* y no de un rechazo o aceptación de *todos* los valores que se encuentran dentro de la ROPE. Se trata *exclusivamente* del parámetro θ de interés. Kruschke (2015b, p.347, Fig. 12.4) utiliza los HDI y la ROPE para criticar la posible no-especificidad de los factores de Bayes. En su ejemplo de monedas, Kruschke toma como base el caso binomial con un 50% de cabezas frente a colas en cada caso, lo que es coherente con el supuesto de la hipótesis nula, que asume $\theta = 0.5$. En primer lugar, hay datos sobre $N = 2$ ensayos con cara una vez.

```
# Bayes Factor with poor precision
# Kruschke p.347
source("DBDA2E-utilities.R")
# p.270
pD <- function(si, Ni, a, b) beta(si+a, Ni-si+b) / beta(a,b)
pD.log <- function(si, Ni, a, b)
exp( lbeta(si+a, Ni-si+b) - lbeta(a,b) )
pD.null <- function(theta.null, si, Ni)
theta.null^si*(1-theta.null)^(Ni-si)
BF.null <- function(pD.null, pD)
{
BF <- pD / pD.null
names(BF) <- ""
return(BF)
}
# example Binomial Bayes-Factor and low precision
# define 50% prob = null value
theta.null <- 0.5
xaxis <- seq(0,1,length=1000)
# [0]
# Kruschke p.347
# left picture
# data
success <- 1
ntrials <- 2
# prior (Haldane prior)
a1b1.prior <- list("a"=0.01, "b"=0.01)
# likelihood
a1b1.likeli <- bino.ab.lik(si=success, Ni=ntrials)
# posterior
a1b1.post <- bino.ab.post(a1b1.prior[["a"]], a1b1.prior[["b"]], success, ntrials)
# results
a1b1.prior
a1b1.likeli
a1b1.post
# plots
par(oma=c(2,1,3,1), "cex.axis"=1, bty="l", mfrow=c(2,2))
plot(xaxis, dbeta(xaxis, a1b1.prior[["a"]], a1b1.prior[["b"]]),
           xlab="Prior (Beta)", ylab="Density", type="l", col="violetred3", main="")
plot(xaxis, dbeta(xaxis, a1b1.likeli[["a"]], a1b1.likeli[["b"]]),
           xlab="Likelihood (Bernoulli)", ylab="Density", type="l",
           col="violetred3", main="")
plot(xaxis, dbeta(xaxis, a1b1.post[["a"]], a1b1.post[["b"]]),
           xlab="Posterior (Beta)", ylab="Density", type="l",
           col="violetred3", main="")
mtext("Prior - Likelihood - Posterior", outer=TRUE,
           line=-0.5, cex=1.5, side=3)
mtext(paste("a = ",a1b1.prior[["a"]]," | b = ",a1b1.prior[["b"]],
           " | successes = ",success, " | trials = ",ntrials,sep=""),
           outer=TRUE, line=-2, cex=1, side=3)
pD.res <- pD(si=success, Ni=ntrials, a=a1b1.prior[["a"]], b=a1b1.prior[["b"]])
pD.null.res <- pD.null(theta.null, si=success, Ni=ntrials)
BF.null.res <- BF.null(pD.null=pD.null.res, pD=pD.res)
```

Si se toma una Prior de Haldane (Beta(0,01; 0,01), Haldane, 1932; véase también Studer, 1996b), que expresa una ignorancia completa sobre si el éxito o el fracaso son posibles en un experimento en absoluto, el factor de Bayes resultante favorece la hipótesis nula con $BF_{01} = 51$. Pero, ¿significa esto, según Kruschke (2015b, p.347), que el parámetro $\theta = 0.5$? En realidad no, porque la Posterior muestra un HDI del 95% con valores que oscilan entre 0.026 y 0.974 (véase la Fig. 6.43). Después de todo 0.974 - 0.026 = 0.948, es decir, el ≈ 94.8% de todo el espectro posible de θ.Si una cosa está clara, es que en estas condiciones ninguna

está claro, entonces es que en estas condiciones no es posible ninguna afirmación clara sobre θ. Aquí están los resultados:

```
> pD.res
[1] 0.004901961
> pD.null.res
[1] 0.25
> BF.null.res
0.01960784
> # OR
> 1/BF.null.res
51
> # classic
> binom.test(success, ntrials, p=theta.null)
Exact binomial test
data: success and ntrials
number of successes = 1, number of trials = 2, p-value = 1
alternative hypothesis:
true probability of success is not equal to 0.5
95 percent confidence interval:
0.01257912 0.98742088
sample estimates:
probability of success
0.5
> # Kruschke DBDA2E-utilities
> HDIofICDF(qbeta, shape1=a1b1.post[["a"]], shape2=a1b1.post[["b"]])
[1] 0.02567744 0.97432255
```

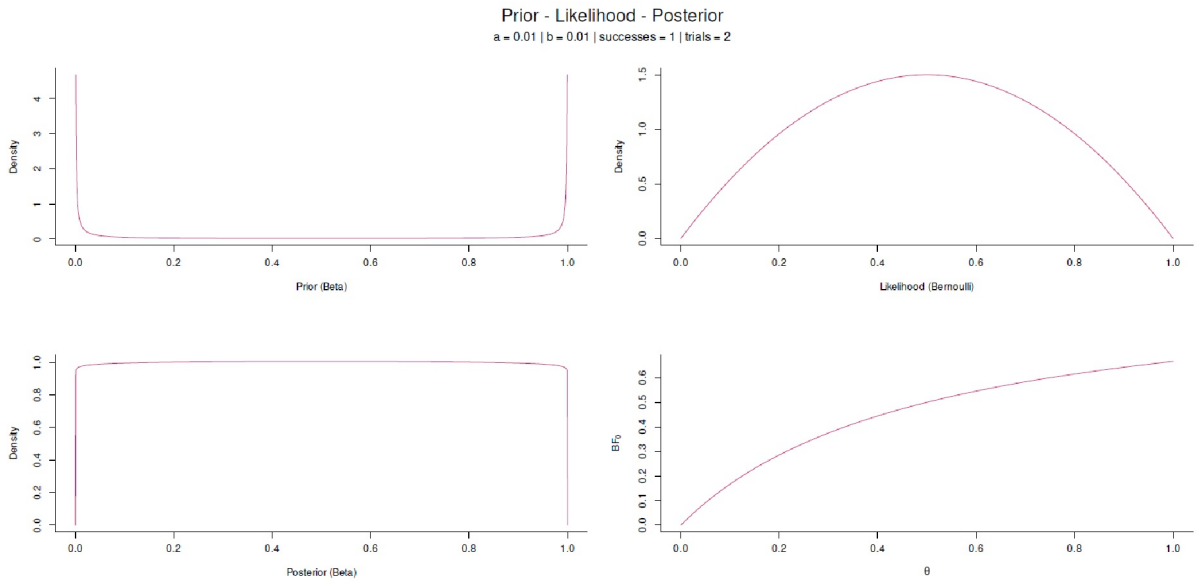


Figura 6.43. *Factores de Bayes y precisión de pruebas de hipotésis nula (caso 1)*

En un segundo ejemplo, hay N = 14 ensayos con 7 veces cara. El resto se repiten como se ha descrito anteriormente (véase la Fig. 6.44).

```
# [1]
# Kruschke p.347
# right picture
# data
success <- 7
ntrials <- 14
```

Si se hace la misma pregunta, la respuesta esta vez también es: *No*. La suposición de que θ = 0.5 dista mucho de ser aceptable. Aunque hay un BF_{01} = 3.14, el HDI al 95% de la Posterior oscila entre 0.266 y 0.734 y sigue cubriendo una gran parte del espectro posible de θ. Aunque la incertidumbre se ha reducido, sigue existiendo incertidumbre, que cubre 0.734 - 0.266 = 0.468, es decir, todavía el 46.8% del espectro total posible de θ. La certeza tiene otro aspecto.

```
> pD.res
[1] 1.942502e-05
> pD.null.res
[1] 6.103516e-05
> BF.null.res
0.3182595
> # OR
> 1/BF.null.res
3.14209
> # Kruschke DBDA2E-utilities
> HDIofICDF(qbeta, shape1=a1b1.post[["a"]], shape2=a1b1.post[["b"]])
[1] 0.2658613 0.7341387
```

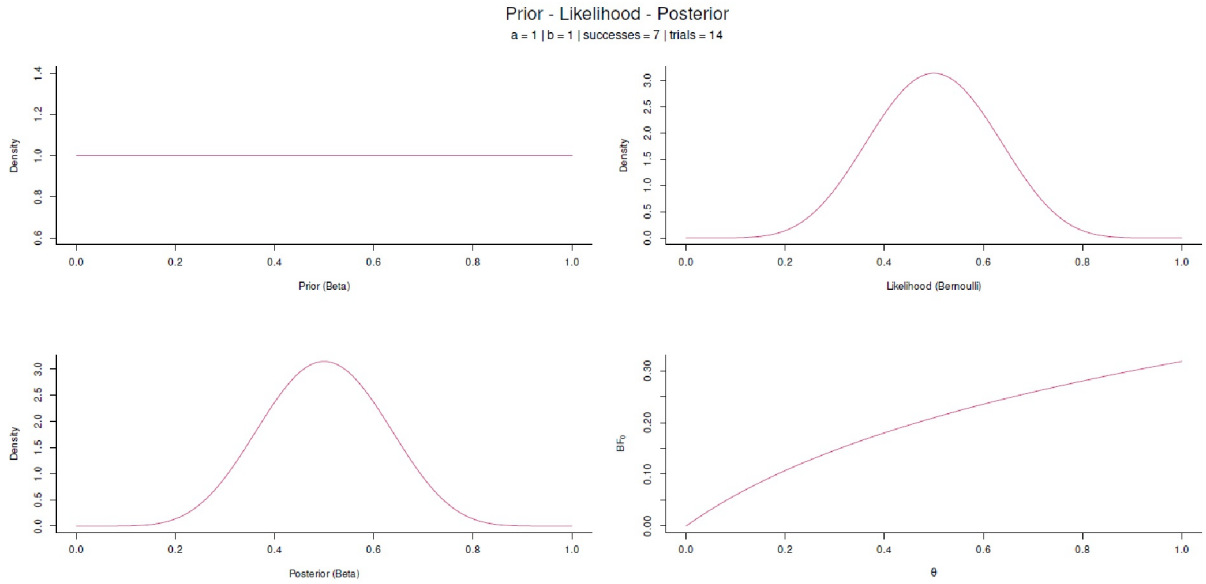


Figura 6.44. *Factores de Bayes y precisión de pruebas de hipotésis nula (caso 2)*

Ahora se pued tomar otra Prior y ver los cambios. Para ello existe la función R `BF.prec.sim()`, que simplemente resume el código R anterior. La siguiente Prior (distribución beta) tiene los parámetros a = 2 y b = 4 y para la Likelihood se eligen s = 7 y n = 24.

```
BF.prec.sim(a=2, b=4, s=7, n=24)
```

Hay muchos casos de uso de ROPE (Kruschke, 2015b, p. 337). Por ejemplo, hay autores (Kruschke, 2018, p.11) que sostienen que la ROPE puede ser una respuesta a la paradoja de Meehl (véase el cap. 4.4.14.3), ya que la comprobación crítica de hipótesis, modelos y, por extensión, más ampliamente, las teorías no debería consistir en rechazar una hipótesis nula selectiva:

> „The concept of ROPE is useful for implementing a solution to a paradox from Meehl (1967, 1997). Theories pursued by null-hypothesis significance testing (NHST) posit merely any non-null effect and are therefore confirmed merely by rejecting the null value of the parameter, regardless of the actual magnitude of the parameter."

Se puede considerar la ROPE como una contrapartida bayesiana de las pruebas de equivalencia clásicas. Partimos de la premisa de que la ROPE no es más que un modelo. En la estadística clásicano hay ningún problema de rechazar prácticamente cualquier hipótesis nula en caso de que el tamaño de muestra crezca cada vez. Esto no significa que un modelo teórico rechazado no pueda hacer un excelente trabajo a nivel práctico (¡incluso como modelo nulo!) o, a la inversa, que un modelo aceptado tenga mucho poder explicativo. La ROPE puede ayudar a limitar drásticamente ese rechazo metodológico-analítico de hipótesis y teorías desplazando la atención del punto al espacio. En términos de proceso, también sería posible hacer que la ROPE fuera cada vez más pequeño en las réplicas para eliminar poco a poco la incertidumbre del sistema, por así decirlo,. y para afinar el trabajo teórico y forzar el examen crítico.

Los límites de la ROPE, al igual que las teorías, no están claramente definidos ni son siempre válidos. Más bien surgen de consideraciones *situacionales prácticas*. Pueden cambiar a través de la replicación o de nueva información. Esto incluye cuando los investigadores quieren probar una teoría de forma mucho más rigurosa, o menos restrictiva, o simplemente aprender algo sobre un área. Esto tiene sentido, porque según el concepto de Lakatos de los programas de investigación, al principio deberían tocarse poco en el núcleo y las pruebas críticas deberían tener lugar en la periferia. Si una teoría se demuestra a sí misma y permite múltiples derivaciones y predicciones, entonces sucesivamente el examen crítico puede intensificarse sucesivamente, un procedimiento que probablemente resulte fatal para el desarrollo de una teoría. Esto es cierto independientemente de lo bien que represente y modele la realidad. Llegados a este punto, cabe mencionar la diferencia entre la pura falsación según Popper o los programas de investigación según Lakatos (véase el capítulo 3).

6.8.4.2.1 *Caso práctico ROPE - Datos de Darwin*. Como ejemplo para ROPE elegimos un conocido conjunto de datos de Charles Darwin (1809-1882), el descubridor de la teoría de la evolución, que se incluye en el paquete R `agridat` como `darwin.maize` (Darwin, 1876, p.16). El conjunto de datos consta de N = 30 observaciones con k = 4 variables cada una. En 1876 Darwin investigó el crecimiento de las plantas de maíz. Las semillas procedían de las mismas plantas madre. Sin embargo, algunas semillas procedían de plantas madre autógamas y otras de plantas madre alógamas. Se plantaron pares de semillas en macetas. Darwin planteó la hipótesis de que las plantas alógamas producían una descendencia más robusta. Según la historia, Darwin no tenía claros los resultados del experimento y ni siquiera su primo Francis Galton pudo ayudarle. Fisher (1935/1973, p.30) afirma entonces que examinó con éxito los datos de Darwin con la prueba *t*. Las cuatro variables del conjunto de datos son maceta (*pot*; 4 niveles), pares (*pairs*; 12 niveles), tipo de fertilización (2 niveles con autofecundación/*self* frente a cruzamiento/*cross*) y la altura medida de las plantas en inches de altura/*height* (`ptII_quan_Bayes_ROPE-BayesFactor.r`).

```
?darwin.maize
darwin.maize
head(darwin.maize)
do.call("rbind",with(darwin.maize, tapply(height, type,
        function(x) c(summary(x),SD=sd(x),VAR=var(x),fivenum2(x))
)))
```

El paquete R `reshape2` permite remodelar los datos para hacerlos más adecuados para el análisis con `melt()` y `dcast()`, que utilizan los autores de `agridat`. En este caso, sin embargo, basta con utilizar `subset()` para extraer los datos correspondientes. A continuación, se establece la prueba *t* bayesiana con `BESTmcmc()` del paquete `BEST` de R (Kruschke, 2013a). En primer lugar, trazamos los datos y las diferencias uno al lado del otro para realizar un análisis exploratorio (véase la Fig. 6.45).

```
do.call("rbind",with(darwin.maize, tapply(height, type,
function(x) c(summary(x),SD=sd(x),VAR=var(x),fivenum2(x)))))
# from ?darwin.maize
dat <- darwin.maize
# Compare self-pollination with cross-pollination
bwplot(height~type, dat, main="darwin.maize")
```

```
dm <- melt(dat)
d2 <- dcast(dm, pot+pair~type)
d2
d2$diff <- d2$cross-d2$self
t.test(d2$diff)
dm
d2
cross <- subset(darwin.maize, type=="cross", select=height)
self <- subset(darwin.maize, type=="self", select=height)
differ <- cross - self
t.test(differ)
# identical
t.test(cross[,1], self[,1], paired=TRUE, var.equal=FALSE)
```

La prueba *t* bayesiana puede almacenarse en un nuevo objeto. Nuestro interés, sin embargo, no son las Posteriors individuales de *cross* y *self*, sino la diferencia de los dos grupos. Se recibe un primer vistazo mediante `plotAll()`.

```
# Bayesian t-test R-Code
darwin.mcmc <- BESTmcmc(d2$cross, d2$self)
plot(darwin.mcmc, credMass=0.87, compVal=2, ROPE=c(-3,2),
       showMode=TRUE, col="grey90", border="white")
pairs(darwin.mcmc)
plotAll(darwin.mcmc)
```

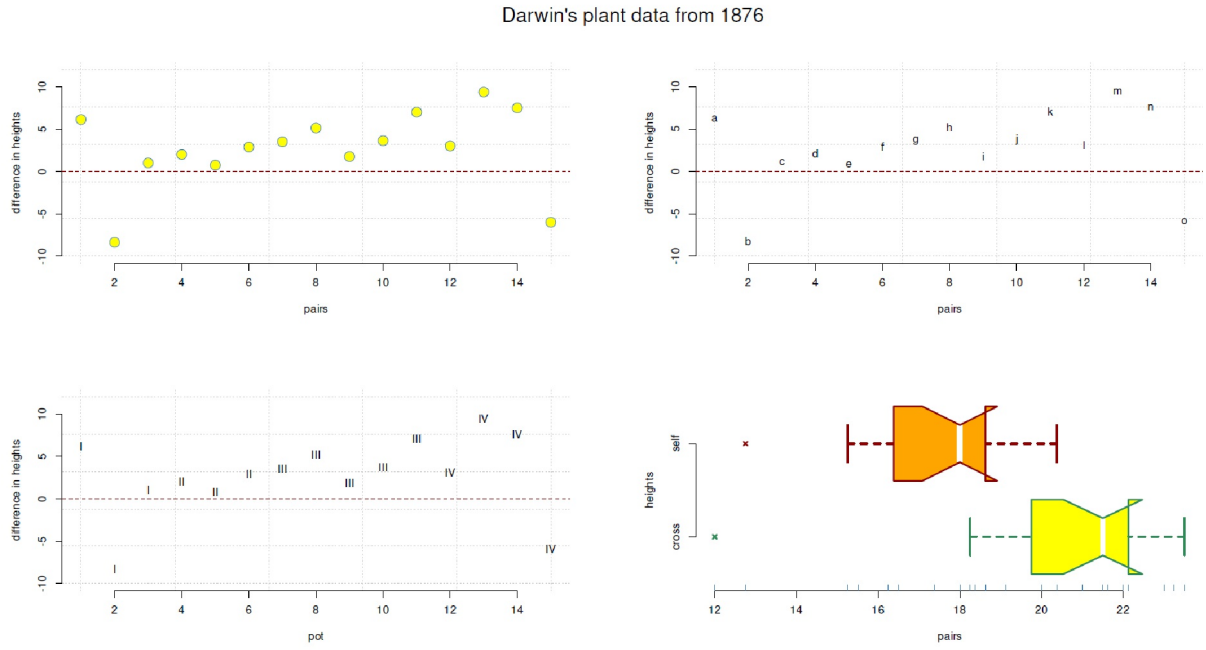


Figura 6.45. *Plantas de maíz – Darwin (datos brutos, varias parcelas).*

Una salida de los resultados de la prueba indica una diferencia sustancial entre las plantas de maíz autógamas y plantas de maíz alógamas.

```
> mean(d2$cross)-mean(d2$self)
[1] 2.616667
> (mean(d2$cross)-mean(d2$self)) / sd(d2$cross)
```

```
[1] 0.7234466
> cohensd(d2$self, d2$cross, sd.theory=sd(d2$cross))
d|mean sd d|pooled sd d|theory sd d corrected|N<50
0.8899077 0.8899077 0.7234466 0.6800262
> # that's our real interest
> mudiff.darwin <- darwin.mcmc$mu1 - darwin.mcmc$mu2
> mean(mudiff.darwin)
[1] 2.988962
> mean(mudiff.darwin > 0)
[1] 0.9916202
> mean(mudiff.darwin > 0.5)
[1] 0.9797104
> mean(mudiff.darwin > 1)
[1] 0.9534609
> mean(mudiff.darwin > 5)
[1] 0.03048939
```

Trazamos la dependencia de la diferencia en la probabilidad de las Posteriores. Un gráfico de ROPE y los predictores posteriores proporciona información más precisa. Elegimos (arbitrariamente) el valor 2.5 como valor comparativo de la diferencia y el ROPE va de 1.4 a 3. En la práctica, por supuesto, estos valores tendrían que fijarse (más) sustancialmente (véase la Fig. 6.46).

```
# plot ROPE and area-in-ROPE R-Code
compvalue <- 2.5
par(oma=c(2,1,3,1), "cex.axis"=1, bty="l", mfrow=c(2,2))
# mcmc difference in means (against zero!, see above)
plotAreaInROPE(mudiff.darwin, credMass=prob, compVal=compvalue,
       maxROPEradius=3, main=expression(paste(mu[diff])))
# raw data
plotAreaInROPE(d2$cross-d2$self, credMass=prob, compVal=compvalue,
       maxROPEradius=10, main="raw")
plotPost(mudiff.darwin, credMass=prob, compVal=compvalue,
       ROPE=c(1.4,3), showMode=TRUE, col="grey90", border="white",
       xlab=expression(theta[2]-theta[1]), ylab="Density",
       main=expression(paste(mu[diff])))
lines(density(mudiff.darwin), col="violetred3", lwd=2, lty=2)
plotPostPred(mcmc1)
mtext("Area in ROPE", outer=TRUE, line=0, cex=1.5, side=3)
```

El análisis MCMC con el paquete coda de R no muestra nada llamativo (no impreso).

```
# MCMC diagnostics R-Code
darwin.post <- as.mcmc(darwin.mcmc)
plot.mcmc(darwin.post, col=c("violetred3","yellowgreen"), bty="n")
```

Para dar una impresión de la distancia y la sostenibilidad del efecto calculamos los Posterior Odds para las distancias d = 0 (ninguna diferencia) hasta d = 10.

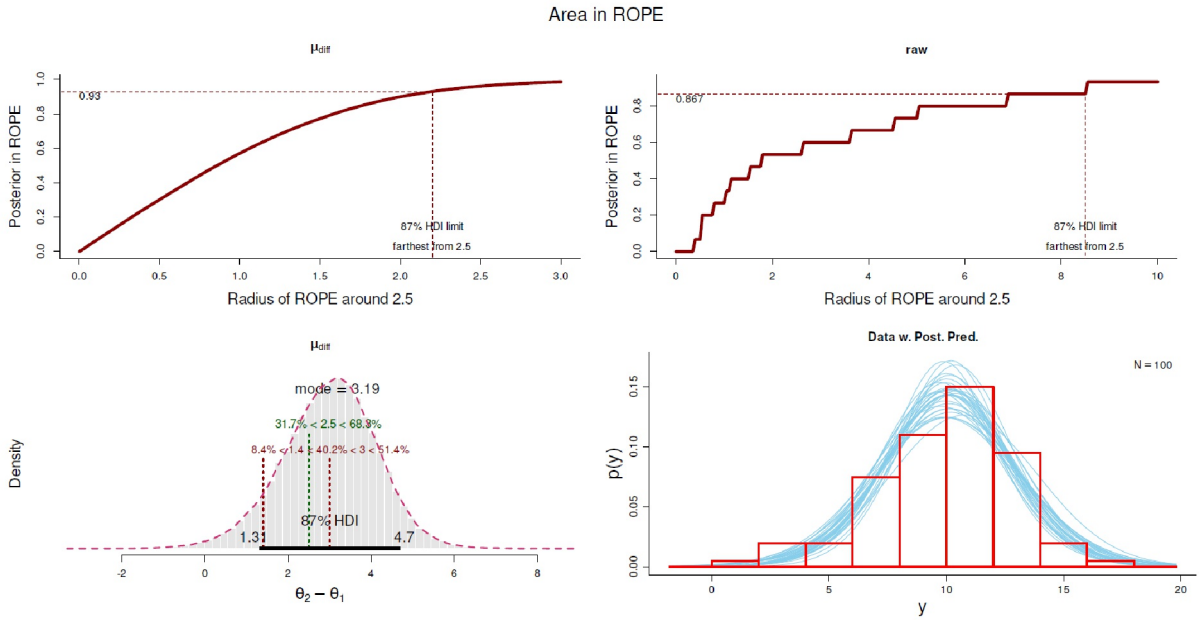


Figura 6.46. *Plantas de maíz de Darwin (diagnóstico del modelo posterior)*

```
# posterior Odds Ratio for comparison values of 'differences' R-Code
compvs <- -2:50
datframe <- data.frame(compv=compvs,OR_post=sapply(compvs,
       function(i) mean(darwin.post > i)/(1-mean(darwin.post > i))))
datframe
head(datframe)
tail(datframe)
xlab <- expression(paste(theta[1]-theta[2]))
ylab <- expression(paste("Odds Ratio"[darwin.post]))
ylab1 <- expression(paste("log(Odds Ratio "[darwin.post],")"))
par(oma=c(2,1,2,1), "cex.axis"=1, bty="l", mfrow=c(2,2))
plot(datframe[-c(1:3),], type="l",col="darkred",bty="n",
       pre.plot=grid(), xlab=xlab, ylab=ylab)
plot(datframe[-c(1:3),1],log(datframe[-c(1:3),2]),
       type="l", col="darkred", bty="n", pre.plot=grid(),
       ylab=ylab1, xlab=xlab)
plot(datframe[-c(1:4),], type="l",col="darkred",bty="n",
       pre.plot=grid(), ylab=ylab, xlab=xlab)
plot(datframe[-c(1:5),], type="l",col="darkred",bty="n",
       pre.plot=grid(), ylab=ylab, xlab=xlab)
mtext("Posterior Odds Ratios", outer=TRUE, line=-1, cex=1.5, side=3)
mtext("for various differences in means", outer=TRUE, line=-2.5, cex=1, side=3)
```

Esto muestra claramente la velocidad con la que las probabilidades posteriores disminuyen al aumentar la distancia. ¿Estamos ahora interesados en una ROPE en torno al rango de $0 < d < 2$ como ejemplo, porque hay una hipótesis dirigida, que es: Las plantas de cereales de fecundación cruzada producen descendencia con mayor crecimiento. Definimos el rango de $0 < d < 2$ por interés de investigación y lo vemos como un rango dentro del cual suponemos que no hay ninguna diferencia real. Los valores fuera de la ROPE se definen así como diferencias existentes. La figura 6.47 muestra la salida correspondiente de la diferencia de las dos distribuciones.

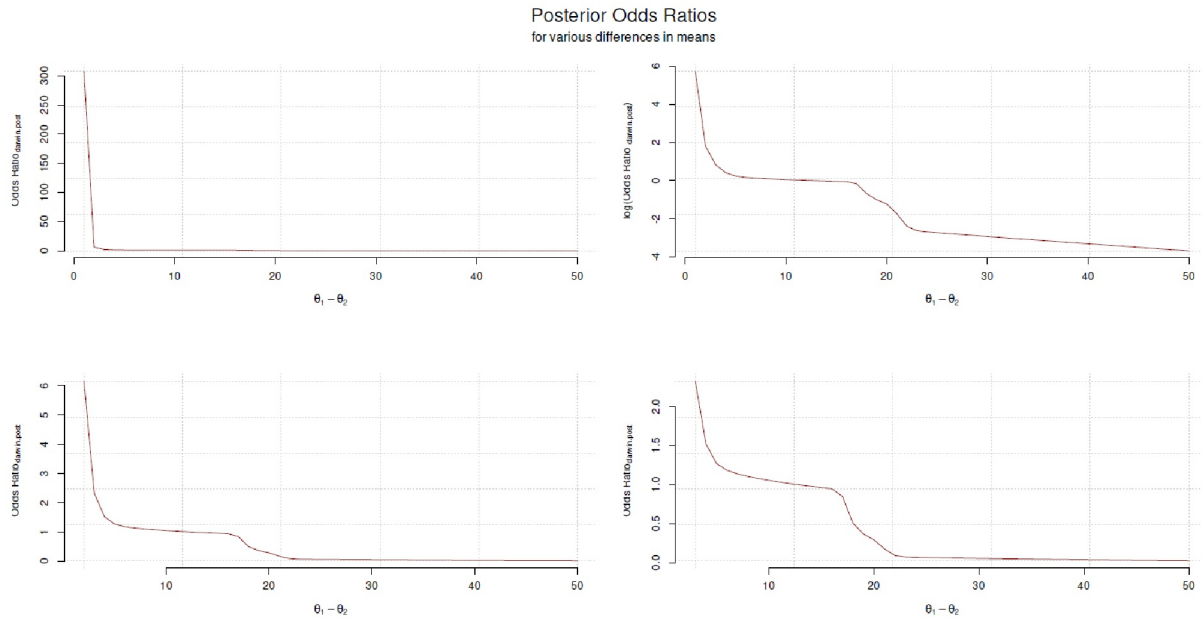


Figura 6.47. *Plantas de maíz de Darwin (probabilidades posteriores frente a diferencias medias)*

```
> 1-mean(mudiff.darwin > 2)
[1] 0.1845763
```

Así, el 18% de las distribuciones posteriores se encuentran dentro de la ROPE y se solapan parcialmente con el HDI del 87%. Aunque se produzca este solapamiento, cabe suponer que las dos distribuciones difieren y lo hacen en la dirección que se ajusta a la hipótesis. Una prueba más cercana sería exactamente cómo de grande resulta ser una estimación robusta de la diferencia, en lugar de si las distribuciones difieren. Esta forma de pensar requiere nuevos datos basados en la replicación y el anclaje en el conocimiento biológico de lo que significa un efecto concreto.

El paquete R `evidence` proporciona un análisis alternativo con el conjunto de datos `darwin`, que se recupera mediante `brm()` del paquete R `brms`, que se basa en `Stan`, para analizar de manera bayesiana las diferencias entre la fecundación de las especies. La llamada es bastante sencilla, pero espera datos distribuidos normalmente:

```
> data(darwin)
> B1Nsir(darwin$difference)
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Bayesian analysis of one Normal sample with SIR-priors
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
sample mean: 21.667 ; sample sd: 38.572
Post. mean: 21.59
95 % cred. int.: -0.279 41.727
Post. std. dev.: 39.478
```

No mucho más sofisticada es `B1Nmean()`, que realiza un análisis bayesiano de una sola muestra, que de nuevo debería seguir el supuesto de una distribución normal. Como Priores se toman valores difusos para la media y la desviación estándar, como puede verse en el código R:

```
# used brms
B1Nmean(darwin$difference, hists=TRUE, pdf=TRUE)
```

– y luego se busca la siguiente línea:

```
prior = student_t(df = 3, location = 0),
        prior_intercept = student_t(df = 3, location = 0)
```

Los resultados corresponden a los gráficos estándar del paquete `brms` de R. `B1Nmean()` es muy rudimentaria y no genera un objeto con el que se pueda seguir trabajando. Por lo tanto, es adecuada para obtener una visión general rápida de una muestra, pero no para análisis planificados y específicos.

> Tarea 6.8: HDI – Cambio de parámetros
>
> La tarea para los lectores interesados en este punto consiste en cambiar las variables para el `HDIcredMass`, el valor de comparación `compVal` y la ROPE y, a continuación, reinterpretar los datos y los resultados de las pruebas en función de las condiciones así modificadas.¿Qué cambia? ¿Qué permanece constante?

6.8.4.3 Calidad de la predicción y estimación del modelo

El debate sobre la *calidad predictiva y la estimación de modelos* de los análisis bayesianos está excelentemente resumido en una entrada de blog de Andrew Gelman (2009a). Damos extractos del mismo para resaltar los puntos clave. Trata de la evaluación de modelos mediante comprobaciones predictivas (= comprobaciones predictivas posteriores).

> "which are then compared to the observed dataset.
> [...]
> – The p-value is the probability that the observed test statistic would be exceeded in replicated data: Pr (T(y.rep) > T(y). [...]
> – The p-value has a uniform (0,1) distribution if the model is true.
> [...]
> All models are wrong, and the purpose of model checking (as I see it) is not to reject a model but rather to understand the ways in which it does not fit the data. From a Bayesian point of view, the posterior distribution is what is being used to summarize inferences, so this is what we want to check. Other people might want to check the model implied by their point estimate"

Gelman,Meng y Stern (1996) destacan la diferencia entre estimación y prueba de modelos. Mientras que la estimación examina las discrepancias entre el modelo y los datos, la comprobación se ocupa de la corrección del modelo. Sin embargo, según los autores, se presta demasiada atención a esta última, en detrimento del ajuste de los modelos, es decir, de la revisión de los modelos ante discrepancias más o menos evidentes entre el modelo y los datos. Mientras que para los problemas más sencillos (por ejemplo, los modelos lineales) existen pruebas convencionales de bondad de ajuste (goodness of fit= GOF, véase también Criterios de información, capítulo 6.8.2), que permiten realizar buenos diagnósticos si se interpretan correctamente, hay muchas preguntas de investigación y análisis para los que no es tan fácil realizar un simple diagnóstico resumido del modelo. Una razón principal es que a menudo faltan distribuciones de

referencia para realizar las pruebas GOF, por ejemplo, en los modelos lineales generales. Los autores (ibíd., p.734) distinguen entre,

> „The classical approach relying on known, or approximately known, reference distributions encounters di culty in at least three kinds of models: models with severe restrictions on the parameters, such as positivity; models with probabilistic constraints, which may arise from a strong prior distribution; and unusual models that cannot be parameterized as generalized linear models."

Debido a la necesidad de poder evaluar modelos complejos, éstos se comparan con datos simulados o, idealmente, con datos replicados. El procedimiento básico de tales evaluaciones se remonta a Box (1980), que habla de "comprobaciones predictivas". El punto de partida es el supuesto, a menudo citado, de que los modelos sólo permiten aproximaciones a una realidad desconocida (exacta). A veces, sin embargo, ni siquiera esto es cierto (Blei, 2011-12-16), por ejemplo cuando los modelos sólo representan formalismos basados en convenciones para especificar determinadas estructuras y, en consecuencia, hipótesis o predicciones. El trabajo básico con modelos incluye los siguientes pasos

1. estimación del modelo
2. evaluación de la discrepancia entre el modelo y los datos
3. revisión

No se trata explícitamente de comparaciones entre modelos, sino de ajustar un modelo y sus parámetros a la información disponible para lograr un *compromiso entre el subajuste y el sobreajuste* (véase el capítulo 6.8.3).

En primer lugar, se estima un modelo bayesiano completo *M* del vector de parámetros θ desconocidos del modelo para los datos observados *y*, de forma que surja una distribución posterior para θ. En el caso de que sea improbable que los datos hayan sido generados por el modelo, éste no se ajustaría y, en consecuencia, las conclusiones basadas en él serían precisas pero incorrectas (Gelman, Meng &Stern, 1996). Según los autores, esto no se puede inferir por la propia distribución posterior, sino por el *proceso frecuentista* de comparar los datos empíricos con una distribución de referencia. Según Box (1980), las ideas frecuentistas y bayesianas están entrelazadas. En este proceso, la parte frecuentista sirve para criticar el modelo de Bayes, mientras que el enfoque bayesiano valora el propio modelo. En realidad, este procedimiento pertenece al ámbito de los métodos mixtos.

En concreto, se generan datos simulados y^{rep} a partir de la distribución posterior manteniendo constante el modelo *M* (véase también bootstrap o prueba de aleatorización en la estadística frecuentista, apartado 4.3.5) y se comparan con los datos empíricos *y*. Para las pruebas críticas, los datos empíricos *y* se localizan en la distribución simulada (Rubin, 1984; Gelman, Meng & Stern, 1996). En realidad, lo que se hace en la simulación es generar una réplica simulada del proceso que se supone que ha generado los datos, es decir, el modelo en cuestión estimado según Bayes *M*. Estas comprobaciones están condicionadas a los datos empíricos observados. Del mismo modo, podría examinarse hasta qué punto los datos simulados replicados se ajustan a los datos reales futuros. Para ello es necesaria la replicación.

La decisión puede tomarse de forma comparable a una prueba frecuentista: Si los datos empíricos se sitúan en los rangos extremos de la distribución simulada, las llamadas colas, debe revisarse el modelo, porque entonces los datos no son típicamente esperados del modelo. Si el modelo se ajusta, la simulación debería producir datos muy parecidos a los observados. Por lo tanto, la prueba del modelo gráfico (véase más adelante) es mucho más adecuada que una prueba frecuentista, ya que se examina todo el intervalo y no sólo un coeficiente proporciona información.

A continuación, *y* son los datos observados, *M* el modelo estimado, θ el vector de parámetros desconocidos e y^{rep} los datos simulados, predichos (replicados). *A(y)* a su vez representa los estadísticos auxiliares, es decir, las funciones de los datos que se mantienen constantes en las réplicas (por ejemplo, el tamaño de la muestra). En la aplicación bayesiana, la Posterior $P(\theta|M,y)$ representa todas las conclusiones

para θ, donde *M* ya contiene la Prior *P(θ)*. La distribución de referencia para las observaciones predictivas y^{rep} (dada *A(y)*) se denomina *distribución predictiva posterior*. Aquí, la distribución simulada posterior de y^{rep} multiplicado por la posterior de θ se integra sobre θ , es decir, $P_A(y^{rep}|M,y)$. y^{rep} es condicional al modelo *M* (en lugar de *M* también se toma en algunas publicaciones como *H* = hipótesis, que es equivalente) y los datos *y*. La replicación *PA* tiene la distribución

$$P_A(y^{\text{rep}} \mid M, \theta) = P\left[y^{\text{rep}} \mid M, \theta, A(y^{\text{rep}})\right] \tag{6.77}$$

La distribución predictiva posterior es (Gelman, Carlin, Stern & Rubin, 2004, p.161; Gelman, Meng & Stern, 1996, p.737)

$$P_A(y^{\text{rep}} \mid M, y) = \int P_A\left[y^{\text{rep}} \mid M, \theta\right) \cdot P(\theta \mid M, y] \, \mathrm{d}\theta \tag{6.78}$$

mientras que el valor *p* clásico como $p(D|H_0)$ esta basado en la estadística de prueba *T*:

$$p_{\text{klassisch}}(y, \theta) = P_A\left[T(y^{rep}) \geq T(y) \mid M, \theta\right] \tag{6.79}$$

Un valor p cercano a cero indica una falta de ajuste del modelo. Sin embargo, lo relevante no es el valor *p*, sino la localización de *T(y)* en la distribución simulada $T(y^{rep})$ – ¿cuánto se ajustan los datos empíricos a los datos simulados? El *valor p predictivo posterior* se define como la probabilidad de que los datos replicados sean iguales o mayores (= más extremos) que los datos observados. La probabilidad resultante del área de la cola (*resultant tail-area probability* = probabilidad en los extremos de la distribución) es

$$\begin{aligned} P_B(y) &= P_A\left[T(y^{\text{rep}}) \geq T(y) \mid M, y\right] \\ &= \int \left[p_c(y, \theta) \cdot P(\theta \mid M, y)\right] \mathrm{d}\theta \end{aligned} \tag{6.80}$$

Esto corresponde al valor *p* clásico promediado sobre la posterior de θ . Este valor se denomina (Rubin, 1984) *valor p predictivo posterior* (Gelman, 2003, 2013d). Esto contrasta con Box (1980), que calcula el valor no sobre la Posterior sino sobre la Prior, que entonces se denomina *p-valor predictivo a priori*. Gelman (2013d) rechaza esto, ya que para él la Prior es sólo una indicación inicial. Box (1980), sin embargo, trata la Prior como una *verdadera distribución* poblacional.

La relación entre la distribución muestral *T(y)* y la distribución simulada $T(y^{rep})$ es idéntica si *T* es una *cantidad pivotal*, condicional a *A(y)* y bajo el modelo *M*. Esto significa simplemente que el modelo estimado a partir de los datos empíricos y las condiciones generales (tamaño de la muestra, etc.) se mantienen completamente constantes, de modo que sólo se sustituyen los datos empíricos por los datos simulados y se obtiene así una nueva distribución. Esto se utiliza para compararla con los datos empíricos según los estadísticos de prueba anteriores y evaluar la calidad. Se dan los datos empíricos *y* y los datos simulados y^{rep} con θ. Estos últimos son independientes entre sí y representan las posibles salidas (= conjuntos de datos) del modelo *M* y los valores de θ.

Gelman, Meng y Stern (1996) amplían este procedimiento mediante el estadístico de prueba *T* para examinar directamente la discrepancia entre los datos y el modelo. Aquí surge claramente la distinción ya introducida entre *prueba y estimación* (= discrepancia). No se trata de la corrección del modelo, sino de

comprender los puntos en los que el modelo y los datos no coinciden. La distribución de referencia de la discrepancia *D* se basa en $P_A(y^{rep} \mid M,y)$ anterior, que es es la distribución marginal de *D*.

$$P_A(y^{\text{rep}}, \theta \mid M, y) = P_A[y^{\text{rep}} \mid M, \theta) \cdot P(\theta \mid M, y] \, \mathrm{d}\theta \qquad (6.81)$$

$P_A(y^{rep}, \theta \mid M, y)$ es ahora la distribución posterior conjunta de y^{rep} y θ. La discrepancia investigada se aplica al par (y^{rep}; θ) y no sólo a y^{rep} como más arriba. Aquí también existe una *probabilidad de cola* $p_{Db}(y)$ para *D* bajo su distribución posterior de referencia. Contiene el caso anterior $p_b(y)$ como caso especial.

$$P_{Db}(y) = P_A[D(y^{\text{rep}}; \theta) \geq D(y; \theta) \mid M, y] \qquad (6.82)$$

Como contrapartida a estos *p*-valores, existen los *u*-valores, que son *p-valores calibrados*. Los *u*-valores son un caso especial cuando dado θ, *T(y)* es casi incidental (es decir, una medida auxiliar independiente de θ) y el *p*-valor posterior $p(T(y^{rep}) > T(y) \mid y)$ tiene una distribución aproximadamente uniforme cuando el modelo es *Verdadero*. Un valor *u* es cualquier función de los datos *y* que tiene una distribución muestral uniforme (Gelman, 2003, p.6). Gelman (2013d, p.2597, cursiva en el original) señala,

> „Under these conditions, p-values less than 0.1 occur 10% of the time, p-values less than 0.05 occur 5% of the time, and so forth. [...]
>
> To clarify, a u-value is any function of the data *y* that has a U(0, 1) sampling distribution. A u-value can be averaged over the distribution of θ to give it a Bayesian flavor, but it is fundamentally not Bayesian, in that it cannot necessarily be interpreted as a posterior probability [...]
>
> In contrast, the posterior predictive p-value is such a probability statement, conditional on the model and data, about what might be expected in future replications.
>
> The p-value is to the u-value as the posterior interval is to the confidence interval."

En principio, en lugar de los valores *p posteriores*, se puede hacer lo mismo con las distribuciones a priori si los datos replicados y^{rep} se generan a partir de la distribución a priori y no a partir de la distribución a posteriori (Gelman, 2003):

$$p(y^{\text{rep}}) = \int p(y^{\text{rep}} \mid \theta) \cdot p(\theta) \, \mathrm{d}\theta \qquad (6.83)$$

La desventaja es que estas comprobaciones predictivas a priori abarcan todo el modelo e incluyen también valores que en realidad han sido eliminados por la Posterior recogida de datos en forma de autoselección y, por tanto, la posibilidad no se ajusta necesariamente a la realidad. Por lo tanto, las Priores pueden entenderse como un caso especial de las comprobaciones predictivas posteriores (Gelman, 2003, S.7):

> „In a posterior predictive check, we are generalizing to future data generated from the same parameter θ, whereas in a prior predictive check, θ^{rep} is redrawn from the model".

Hay argumentos en contra de este enfoque. Kruschke (2013c) argumenta que está de acuerdo con Gelman y sus colegas en que los resultados bayesianos deberían o incluso deben someterse a pruebas de falsación. Sin embargo, esto debería hacerse con métodos bayesianos y no recurriendo a los problemáticos *p*-valores de la estadística clásica introducidos de nuevo por la puerta de atrás. Kruschke (2013c) utiliza dos estudios de casos para ilustrar el problema y subraya que las comprobaciones predictivas a posteriori también deberían realizarse bayesianamente. Al hacerlo, defiende la ampliación del modelo básico frente a uno reducido, cuando los dos modelos están anidados. En caso de información a priori idéntica, pueden tomarse factores de Bayes para contrastar modelos diferentes. Si la información a priori no es comparable, la

comparación es más difícil porque las causas de las diferencias observables sólo pueden atribuirse a causas claras hasta cierto punto y existe un entrelazamiento irresoluble de la información a priori y el modelo. Esto, a su vez, es un problema de diseño y no matemático, y debe resolverse en consecuencia a nivel de diseño. En un caso de uso, en cambio, se pueden hacer predicciones y contrastarlas con los acontecimientos futuros para averiguar la superioridad del modelo respectivo. Pero incluso en este caso existe cierta incertidumbre en el caso de información a priori diferente. Una extensión sería en el sentido de los métodos mixtos (capítulo 13) comparar diferentes niveles de análisis en cuanto a su contenido y no puramente a nivel matemático – por ejemplo, recomendaciones prácticas para la elección de una intervención o similar. Esto tendría la ventaja de que los distintos niveles de análisis, a pesar de sus diferencias, podrían conducir a conclusiones o decisiones convergentes a nivel sustantivo-práctico.

Además de todas estas posibilidades, siempre existe la necesidad de verificar gráficamente los modelos. ya sean bayesianos o frecuentistas clásicos.

6.8.4.4 Tasas de aprobación de muestras de casos en drogodependencias en régimen de hospitalización.

A modo de ejemplo, mostramos el procedimiento para las revisiones predictivas posteriores. Nos guiamos por varios procedimientos de entradas de blog de Kruschke (2012a, 2016c, 2017b) y Nalborczyk (2018). El conjunto de datos utilizado son las tasas de aprobados de *start again*, que se describen con más detalle en el capítulo 6.15.2. El conjunto de datos abarca 26 años, de 1992 a 2017, con N = 602 clientes (x = 23.15, s = 8.46). Para los 298 seguimientos, (anualmente) x = 11.46 y s = 5.43, lo que arroja una tasa de aprobados de 298/602 = 0.495, es decir, casi el 50%. También podríamos tomar otra, por ejemplo una beta informada (1.79; 3.068), que asume que el fracaso y el éxito son posibles, pero se centra en un éxito < 50%. Elegimos entonces una mediana del 35%. Los límites inferior y superior de la masa del 90% de la distribución a priori se fijarían en el 15% y el 65%, respectivamente. Determinamos los parámetros beta a y b con `beta.determine.opt()` utilizando la minimización del error cuadrático (véanse las explicaciones en el capítulo 6.12). Este procedimiento se desvía del procedimiento de Studer (1996b, 1998) descrito en el capítulo 6.15.2. Como se verá más adelante (véase el capítulo 6.15.2), dada la cantidad de datos (influencia de la Likelihood), hay sin embargo muy pocos cambios en la Posterior. Corresponde a los lectores reproducir este procedimiento.

A partir de los datos empíricos y la probabilidad a priori, la probabilidad a posteriori puede determinarse analíticamente con la $mediana_{post}$ = 0.495, μ_{post} = 0.495 y σ_{post} = 0.02, así como los parámetros beta a_{post} = 299 y b_{post} = 305, ya que la distribución beta a priori es conjugada con la Likelihood y la probabilidad a posteriori es también una distribución beta. El mapeo de estos parámetros en una cuadrícula da como resultado valores posteriores para el rango válido de valores de cero a uno. En la Figura 6.48 se muestran los valores a priori, de Likelihood y a posteriori. Como se puede ver, la Likelihood domina casi por completo a la Posterior. Aquí sigue el código R (`ptII_quan_Bayes_PPC_model-check.r`):

```
# read in 'sa.all' and 'sa.bino' R-Code
sa.all <- read.table("startagain_statistics_1992-2017_all-out.tab",
        sep="\t", header=TRUE)
sa.all.d <- dim(sa.all)
sa.bino <- read.table("startagain_statistics_1992-2017_bino.tab",
s       ep="\t", header=TRUE)
sa.bino.d <- dim(sa.bino)
# plot results of prior likelihood, and posterior for the year 2017
thetas <- seq(0,1,0.01)
par(oma=c(2,1,2,1), "cex.axis"=1, bty="l")
plot(thetas,
        dbeta(thetas,shape1=sa.bino[sa.all$year == 2017,"a.lik"],
        shape2=sa.bino[sa.all$year == 2017,"b.lik"]),type="l",
        pre.plot=grid(), col="yellowgreen", lty=2, lwd=2,
        ylab="Density", xlab=expression(theta))
lines(thetas,dbeta(thetas,shape1=1,shape2=1),
        lty=2, lwd=2, col="blue")
lines(thetas, col="darkred", lwd=1, lty=1,
```

```
        dbeta(thetas,shape1=sa.bino[sa.all$year == 2017,"a.post"],
        shape2=sa.bino[sa.all$year == 2017,"b.post"]))
mtext("Prior, Likelihood and Posterior", outer=TRUE,
        line=-0.5, cex=1.5, side=3)
mtext("success rates start again 1992-2017", outer=TRUE,
        line=-2, cex=1, side=3)
par(fig=c(0,1,0,1), oma=c(1,0,0,0), mar=c(0,0,0,0), new=TRUE)
plot(1, type="n", bty="n", xaxt="n", yaxt="n")
categs <- c("prior","likelihood","posterior")
        legend("bottom", legend=categs, lty=c(2,2,1), lwd=c(2,2,1),
        xpd=TRUE, horiz=TRUE,
        col=c("yellowgreen","blue","darkred"), bty="n", cex=.9)
```

A continuación, se simulan mil valores aleatorios a partir de la posterior mediante `rbeta()`,

```
# number of replicated samples
nsims <- 1e3
z <- sa.all[sa.all.d[1],"s"]
a <- sa.bino[sa.bino.d[1],"a.prior"] + z
b <- sa.bino[sa.bino.d[1],"b.prior"] + z
n <- sa.all[sa.all.d[1],"N.cs"]
# generating nsims theta values from posterior
# ie. random values from posterior distribution with
# a.post = a.prior + z
# b.post = b.prior + n - z
# z = si = successes
# thetas <- rbeta(nsims, a + z, b + n - z)
year <- 2017
a.post <- sa.bino[sa.all[,"year"] == year,"a.post"]
b.post <- sa.bino[sa.all[,"year"] == year,"b.post"]
a.post
b.post
seed <- 0987
set.seed(seed)
thetas <- rbeta(nsims, shape1=299, shape2=305)
head(thetas)
tail(thetas)
# length of the vector to replicate for prediction
N.cs <- sa.all[sa.all[,"year"] == year,"N.cs"]
N.cs
# number of successes to have something to compare
s.cs <- sa.all[sa.all[,"year"] == year,"s.cs"]
s.cs
s.cs/N.cs
```

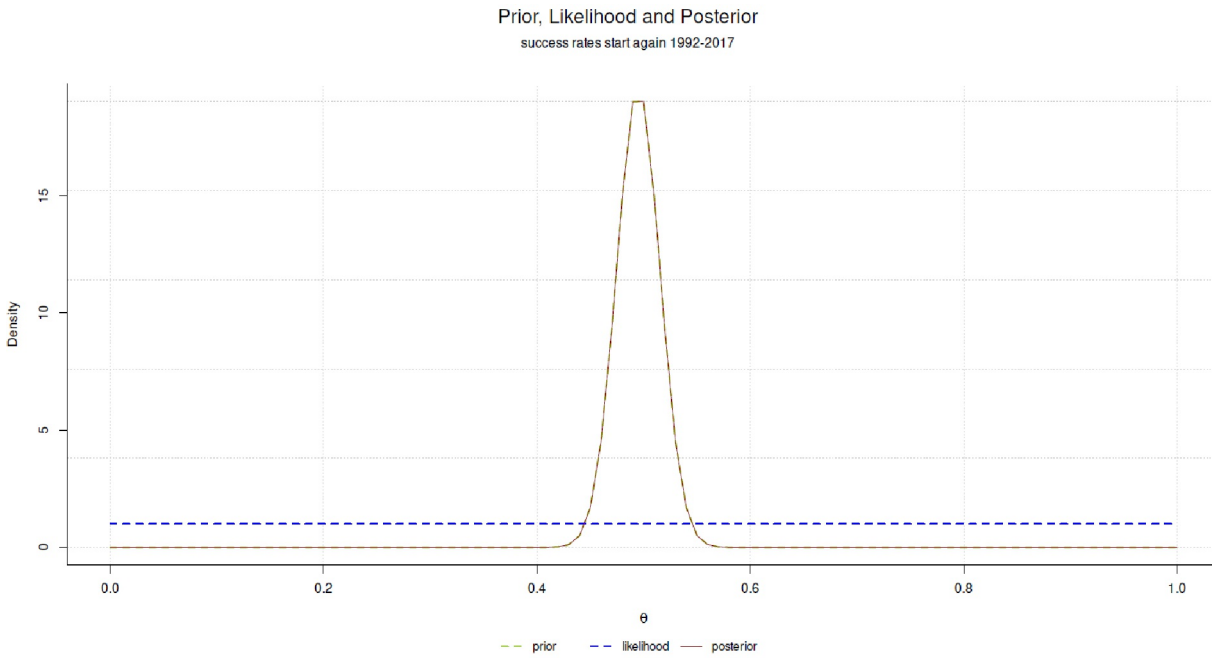


Figura 6.48. *Start again* (*porcentajes de aprobados utilizando Prior, Likelihood y Posterior)*

y junto con las probabilidades de los valores simulados a partir de la Posterior, se generan los valores predictivos y^{rep}. Aquí hay que decidir cuántos en cada caso. Por un lado, se podrían generar valores para todo el periodo investigado o sólo para el año siguiente a lo largo de una estimación robusta de la ocupación de los clientes, es decir, que no se sobrestimen las fluctuaciones locales a corto plazo. Nosotros optamos por todo el periodo. Lo que se elija exactamente depende del interés de la investigación. La generación de la y^{rep} se realiza de tal manera que para cada una de las θ*s* extraídas de la posterior anterior, se genera una muestra aleatoria a lo largo del tamaño de muestra objetivo. El tamaño de la muestra corresponde a la clientela total durante el período 1992-2017, es decir, N = 602, produciendo valores de cero (= ninguna ejecución/fracaso) o uno (= ejecución/éxito). La probabilidad de *éxito* frente a la de *fracaso* se basa en la respectiva θ de la Posterior. Esto indica la probabilidad *p* para el *éxito*, mientras que el elemento complementario se aplica para el *fracaso*, es decir, q = 1 - *p*. En R esto se puede implementar con `sample()`.

```
# length of the vector to replicate for prediction
N.cs <- sa.all[sa.all[,"year"] == year,"N.cs"]
N.cs
# number of successes to have something to compare
s.cs <- sa.all[sa.all[,"year"] == year,"s.cs"]
s.cs
s.cs/N.cs

# adopted to our needs...
# draw from posterior to predict
Yrep <- sapply(1:length(thetas),
       function(i) sample(c(0,1), N.cs, replace=TRUE,
       prob=c(thetas[i], 1-thetas[i]))
)
str(Yrep)
head(Yrep)
table(Yrep)
```

Se crea una gran matriz

```
> dim(Yrep)
[1] 602 1000
```

donde las filas representan θ y las columnas los elementos de la muestra. Si se forma las sumas sobre las filas, el resultado es un valor de *éxito* por θ, que puede convertirse en una probabilidad esperada por θ (porcentaje de éxito simulado esperado) calculando el valor medio por simulación.

```
# success rate for each replication/ simulation of predictive posterior
Trep <- apply(Yrep, 2, function(x) sum(x)/N.cs)
head(Trep)
tail(Trep)
describes(Trep)
```

Se pueden trazar estas tasas de éxito predictivo desde BEST con `plotPost()` y se marcan con una línea vertical la moda posterior (= MAP) de los valores de 1992-2017 (véase la Fig. 6.49).

```
# description of predictive posterior
data.frame(T_rep=round(c(summary(Trep), SD=sd(Trep), VAR=var(Trep),
       fivenum2(Trep)),3))
# original values
sa.bino[sa.all[,"year"] == year,]
# comparison value
Ty <- sa.bino[sa.all[,"year"] == year,"mode.post"]
Ty
credMass <- 0.95
ROPE <- c(Ty-0.08,Ty+0.08)
# plot predictive distribution
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l")
plotPost(Trep, compVal=Ty, breaks=20, credMass=credMass, ROPE=ROPE,
       col="#E6E6E6", xlab=expression(T(y^rep) ) , ylab="Density")
lines(density(Trep), col="darkred", lwd=2, lty=2)
MAP.post <- sa.bino[sa.all[,"year"] == year,"mode.post"]
abline(v=MAP.post, col="orange",lwd=2,lty=2)
legend("topright",
       legend=paste("MAP 1992-2017\n[p = ",signif(MAP.post,3),
       "]",sep=""),
       lty=2, lwd=2, col="orange", bty="n", cex=.9)
mtext("Simulating Posterior for replications of Y", outer=TRUE,
       line=-2, cex=1.5, side=3)
```

Comparamos los valores originales *Ty* con las predicciones `Trep` y obtenemos el porcentaje de la masa de la distribución por comparación:

```
> # how many Trep > Ty ?
> # ie. are there more switches in
> # Trep = posterior predictive compared to Ty = empirical?
> # ie. location of T_y in the distribution of T_rep
> sum(Trep > Ty)
[1] 629
> sum(Trep > Ty)/nsims
[1] 0.629
> sum(Trep > Ty)/length(thetas)
[1] 0.629
> # probability
> mean(Trep > Ty)
[1] 0.629
> 1/mean(Trep > Ty)
[1] 1.589825
```

```
> mean(Trep - Ty > 0)
[1] 0.629
> # inverse= Bayesian p-value (p < alphacrit)
> 1-mean(Trep - Ty > 0)
[1] 0.371
> 1-mean(Trep - Ty > 0.05)
[1] 0.928
> 1-mean(Trep - Ty > 0.1)
[1] 0.998
> 1-mean(Trep - Ty > 0.2)
[1] 1
```

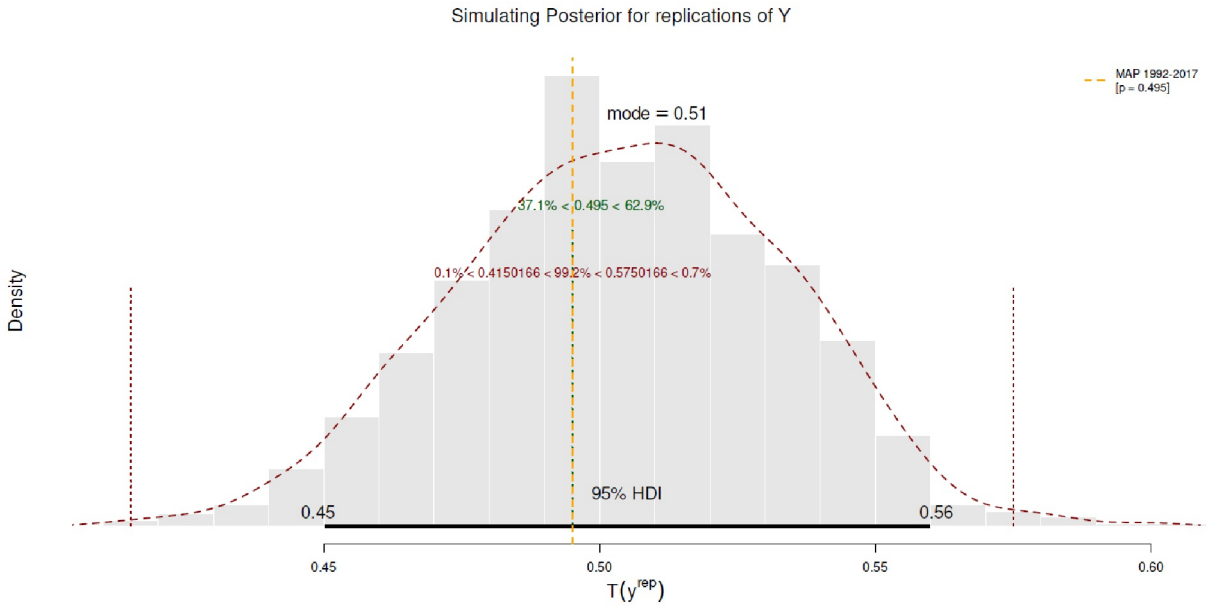


Figura 6.49. *Start again (porcentajes de aprobados, distribución predictiva)*

A continuación se calcula la base del valor *p* bayesiano observando qué porcentaje de la masa de la distribución (diferencia Trep - Ty) se encuentra por encima de un umbral crítico. El rango de estos umbrales va de 0.001 a 0.15.

```
sekstart <- -0.10
sek <- seq(sekstart,0.15,0.001)
postps <- vector()
for(i in 1:length(sek)) postps[i] <- mean(Trep-Ty > sek[i])
postps.tab <- data.frame(crit=sek,p=postps)
head(postps.tab)
tail(postps.tab)
```

Lo representamos gráficamente (véase la Fig. 6.50)

```
# plot
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l")
plot(sek,postps, bty="n", type="l", col="darkred", pre.plot=grid(),
       xlab="Difference = critical comparison value", ylab="p > crit", main="")
mtext("Bayesian p-value [ = 1 - p]", outer=TRUE, line=-2, cex=1.5, side=3)
# zero line
plot.comp.bpv(sek,postps,compcrit=0)
# empirical posterior mode
```

```
plot.comp.bpv(sek,postps,compcrit=mean(Trep)-Ty, Trep=Trep, colo="red")
# one sided hypotheses:
# some interesting value -> very high Bayesian p-value
plot.comp.bpv(sek,postps,compcrit=0.06, Trep=Trep, colo="yellowgreen")
# some interesting value -> very low Bayesian p-value
plot.comp.bpv(sek,postps,compcrit=-0.06, Trep=Trep, colo="yellowgreen")
```

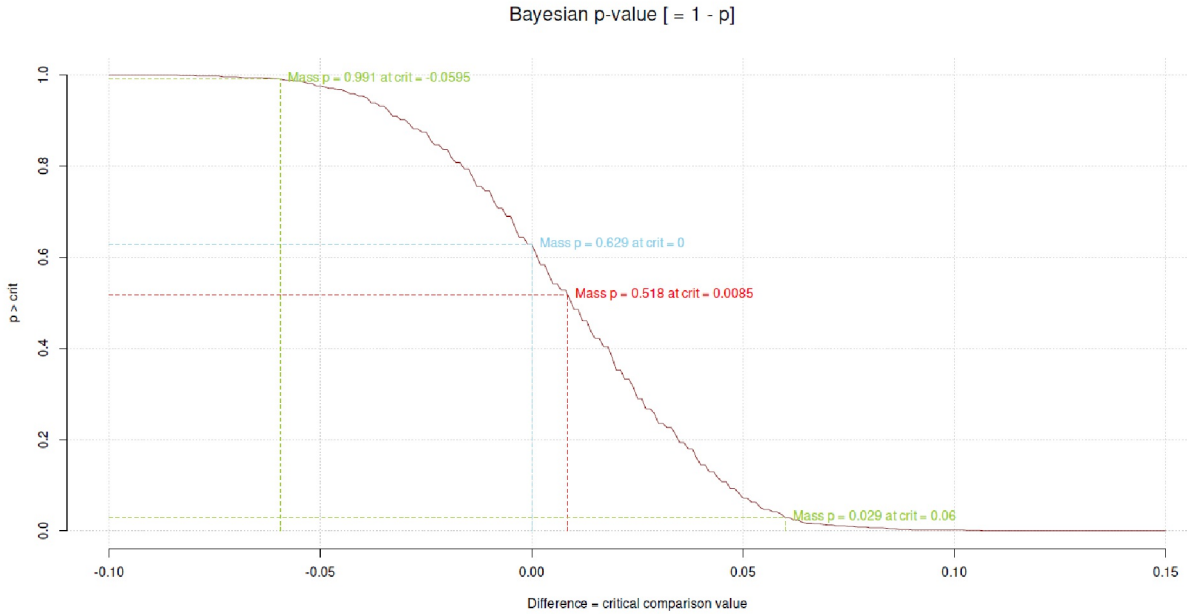


Figura 6.50. *Start again (porcentajes de aprobados, p-valor bayesiano)*

y comparamos el vector `postps` con un valor crítico empírico `Ty`:

```
# compare to crit
postps[postps > Ty]
postps[postps < Ty]
```

A partir de aquí, puede obtenerse el valor *p* bayesiano. Debe tenerse en cuenta que este valor es frecuentista, es decir, dice algo sobre la probabilidad de los datos empíricos en el contexto de la simulada distribución predictiva (simulada de los porcentajes de éxito). En nuestro caso, estamos comparando la diferencia de los valores predictivos con los esperados, por ejemplo, la moda posterior, para diferentes valores de diferencia, para hacernos una mejor idea del entorno "como de costumbre" más que para tomar decisiones fijas de sí o no.

```
> # Or we can compute a Bayesian p-value as ...
> # see Gelman et al., 2013, page 146:
> # pB=Pr(T(yrep,j)   T(y,j)|y)
> # equivalent to sum(Trep <= Ty) / nsims
> 1 - sum(Trep > Ty) / nsims
[1] 0.371
> sum(Trep <= Ty) / nsims
[1] 0.371
> mean(Trep > Ty)
[1] 0.629
> 1-mean(Trep > Ty)
[1] 0.371
> # e.g. =0.0079 ~ 0.008 -> ie. "p" < < 0.05 / 0.01 etc.
> # if interpreted as a p-value
> # = prob whether observations = y = data are probable
```

```
> # under this type of model (= hypothesis)
> # with the parameters chosen
```

Esto muestra que no se producen diferencias estadísticamente significativas dentro de una diferencia de interés de 0.2. Esto puede representarse gráficamente con `plot.comp.bpv()` para explorar el entorno más fácilmente (véase más arriba o la Fig. 6.50). Este valor *p* debe interpretarse *clásicamente* como que no hay pruebas de que el modelo sea inadecuado o que los datos y el modelo no encajan debido a una gran desviación. No se trata de una probabilidad de que el modelo sea "correcto" o "se ajuste", sino que se orienta hacia todas las afirmaciones hechas hasta ahora sobre la estadística clásica y los p-valores. Esta no es la única razón por la que Kruschke (2013c) señala que las comprobaciones predictivas posteriores deben ser bayesianas y no frecuentistas. Se puede realizar un trabajo equivalente con otros modelos. Los modelos se vuelven más complejos, pero el procedimiento básico permanece constante.

6.8.4.5 Evaluación gráfica de modelos

La evaluación gráfica de modelos es muy similar a las explicaciones dadas en el capítulo 5 sobre el análisis exploratorio de datos. Los datos se examinan y exploran gráficamente para comprobar si hay desviaciones respecto a los supuestos y expectativas del modelo (Gabry, Simpson, Vehtari, Betancourt & Gelman, 2019). Gelman y Shalizi (2010, 2013) describen un análisis bayesiano detallado de datos en el que solo a través de la revisión gráfica *se falsificó* el modelo adoptado originalmente para sustituirlo por un nuevo más ajustado. Falsificación no significa aquí descartar un modelo en favor de otro, sino descartar el modelo actual para cambiarlo por otro mejor. En concreto, en el primer paso, los autores crearon un modelo lineal jerárquico (con intercepción variable) para responder a una pregunta de la ciencia política, a saber, la relación entre el comportamiento político de voto y la desigualdad de ingresos en los estados de EE.UU. Sólo cuando este modelo resultó no ajustarse cuando el modelo se consideró visualmente, se amplió el modelo para incluir una pendiente variable. El factor decisivo para esta decisión fue que la indicación podía tomarse de una cifra y no sobre la base de un coeficiente estadístico. Se visualizó "la respuesta media de la encuesta y las curvas ajustadas para las distintas categorías de ingresos dentro de cada estado" (Gelman & Shalizi, 2013, p.13). Los autores (ibíd.) muestran algo similar con otro modelo, en el que se suponía siempre el mismo efecto del tratamiento para los efectos *antes-después*, sin un término de interacción. El análisis gráfico mostró que el modelo no representaba bien los datos porque "La línea de las unidades de control tenía una pendiente mucho más pronunciada que la de las unidades tratadas. Ajustamos un nuevo modelo, y tenía una historia completamente diferente sobre lo que significaban los efectos del tratamiento" (ibíd., p.24). En consecuencia, se añadió al modelo un término de interacción. Además del aspecto falsificador del modelo previamente favorecido, la ilustración podía contribuir inmediatamente a la solución del problema, es decir, en qué dirección había que cambiar el modelo (por ejemplo, la inclusión del término de interacción). Los autores comentan esto con (ibid., p.24f.):

> „This pattern of higher before-after correlation in the control group than the treated group is quite general (Gelman, 2004), but at the time we did this study we discovered it only through the graph of model and data, which falsi ed the original model and motivated us to think of something better. In our experience, falsi cation is about plots and predictive checks, not about Bayes factors or posterior probabilities of candidate models."

Esto es exactamente lo que Tukey (1977) dijo sobre el AED: *comprender y encontrar estructuras*. Observamos el proceso en un modelo casi ideal para hacernos una idea de cómo puede ser casi perfecto, de modo que podamos basarnos en él para comprender mejor los modelos más débiles. Se crean dos modelos. El primero se parece a una regresión casi perfecta cuyos valores poblacionales especificamos. El segundo modelo utiliza un control de dos pasos en lugar de un control continuo. Para ambos modelos lineales se obtienen los estadísticos habituales y se trazan los modelos para estimar gráficamente de la Posterior (`ptII_quan_Bayes_PPC_model-check-graph.r`).

```
# case 1 - linear regression - almost ideal fit
n <- 100
x <- 1:n
set.seed(887766)
y <- x + rnorm(100,0,10)
plot(x,y)
xymodel <- data.frame(x, y)
```

Todo comienza con una simple línea oblicua con un ángulo de 45° para el predictor *x*. Para la variable dependiente *y*, simplemente añada algo de ruido distribuido aleatoriamente a *x* con *y ~ N(0,10)*, es decir, valores distribuidos normalmente con un valor esperado de cero y una varianza de 10. Las variables se correlacionan prácticamente de forma lineal máxima entre sí, pero sin ser completamente idénticas.

```
> cor(x,y)
[1] 0.9609738
```

Un modelo lineal simple corresponde a las expectativas y lo mismo ocurre con el gráfico correspondiente. (véase la Fig. 6.51).

```
# frequentist analysis
lm.xymodel <- lm(y ~ x, data=xymodel)
summary(lm.xymodel)
display(lm.xymodel)
# plot
plot(xymodel$x, xymodel$y, pch=21, cex=1.1, xlab="x", ylab="y",
       bg="yellow", col="steelblue", pre.plot=grid(), bty="n", main="")
abline(lm.xymodel, lty=2, col="red", lwd=2)
lines(lowess(x,y), lty=1, col="darkgreen", lwd=2)
par(mfrow=c(2,2))
plot(lm.xymodel, col="purple", main="")
```

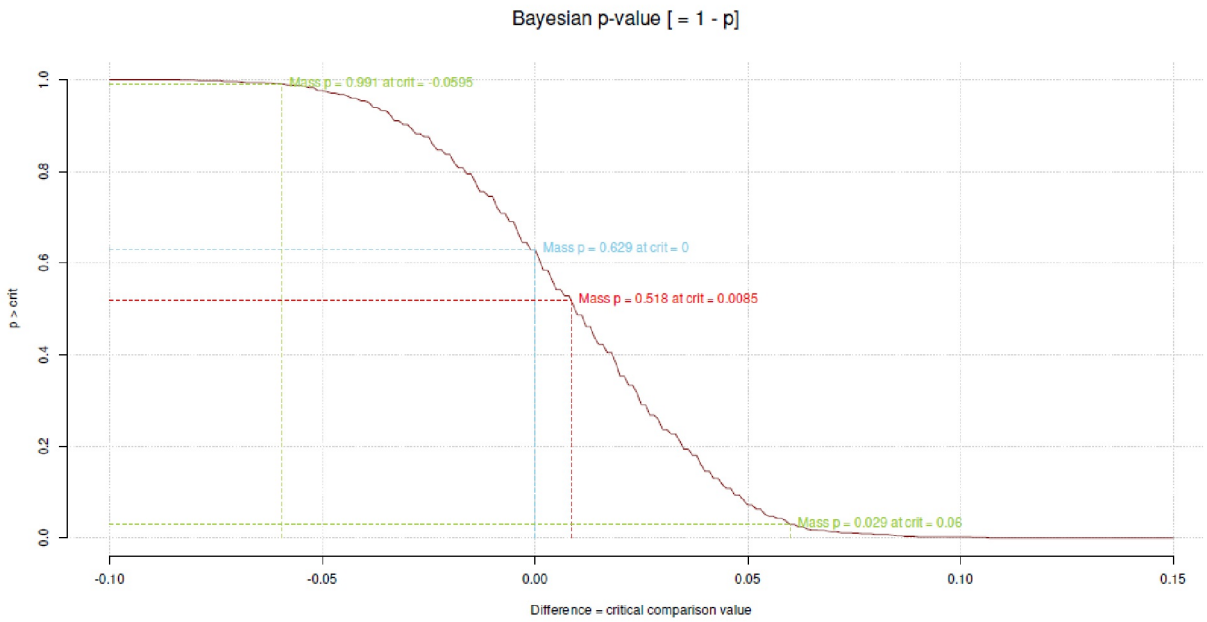


Figura 6.51. *Modelo lineal (ajuste casi perfecto, diagrama de dispersión)*

Ahora seleccionamos para el predictor el intercepto y la varianza Prior, por ejemplo, para el predictor seleccionamos *x ~ N(1, 5)*. El intercepto y la varianza se dan cada uno – en notación R – una Prior a `student_t(3, 50, 29)` y `student_t(3, 0, 29)` respectivamente. El modelo bayesiano se describe con

Stan (2019b) y brm() estimada del paquete brms de R. Para los análisis gráficos recurrimos a el paquete R bayesplot, que hace automáticamente las llamadas de las funciones R en brms.

```
# Bayesian analysis with Stan and brms package R-Code
P <- c(prior(normal(1,5), class="b", coef="x"),
       # prior(student_t(2,0,10), class="Intercept"),
       # prior(student_t(2,0,10), class="sigma"))
prior(student_t(3,50,29), class="Intercept"),
       prior(student_t(3,0,29), class="sigma"))
brm.xymodel <- brm(y ~ x, data=xymodel, family="gaussian",
       save_all_pars=TRUE, prior=P)
```

Se puede examinar el objeto resultante con más detalle con summary(), plot() y pairs() como es habitual en R. Efectos, residuos, predicciones (también sobre nuevos datos), matrices de varianza-covarianza, muestras posteriores y muchas otras cantidades pueden extraerse o generarse fácilmente a partir del modelo. pp_check() genera una simulación de comprobación predictiva posterior y^{rep} en comparación con los datos empíricos y. Para ello, existen varias opciones de trazado disponibles, cuyos detalles se pueden encontrar en la página de ayuda de pp_check(). Para empezar, valen type="ecdf_overlay", type="ecdf_scatter_avg" y type="stat". Gabry, Simpson, Vehtari, Betancourt y Gelman (2019) proporcionan más orientación sobre qué gráficos son útiles para el diagnóstico de modelos.

```
summary(brm.xymodel)
plot(brm.xymodel)
pairs(brm.xymodel)
pp_check.brmsfit(brm.xymodel, nsamples=100)
pp_check.brmsfit(brm.xymodel, nsamples=100,
       type="ecdf_overlay")
pp_check.brmsfit(brm.xymodel, nsamples=100,
       type="error_scatter_avg")
pp_check.brmsfit(brm.xymodel, nsamples=100,
       type="stat")
```

Se pueden comprobar hipótesis. Por ejemplo, podríamos estar interesados en saber si el intercepto es es cero o si la pendiente del predictor es igual a uno.

```
# hypotheses
hypo0.int.greater.zero <- hypothesis(brm.xymodel, "Intercept > 0")
hypo0.int.smaller.zero <- hypothesis(brm.xymodel, "Intercept < 0")
hypo1.b.eq.1 <- hypothesis(brm.xymodel, "x = 1")
```

Examinemos más detenidamente esta última hipótesis en la salida y tracemos las Posteriores correspondientes de esta hipótesis (véase la Fig. 6.52):

```
> hypo1.b.eq.1
Hypothesis Tests for class b:
Hypothesis Estimate Est.Error CI.Lower CI.Upper
1 (x)-(1) = 0 0.07 0.03 0 0.13
Evid.Ratio Post.Prob Star
1 NA NA *
---
'CI': 90%-CI for one-sided and 95%-CI for two-sided hypotheses.
'*': For one-sided hypotheses, the posterior probability
exceeds 95%; for two-sided hypotheses, the value tested against
lies outside the 95%-CI. Posterior probabilities of point
hypotheses assume equal prior probabilities.
```

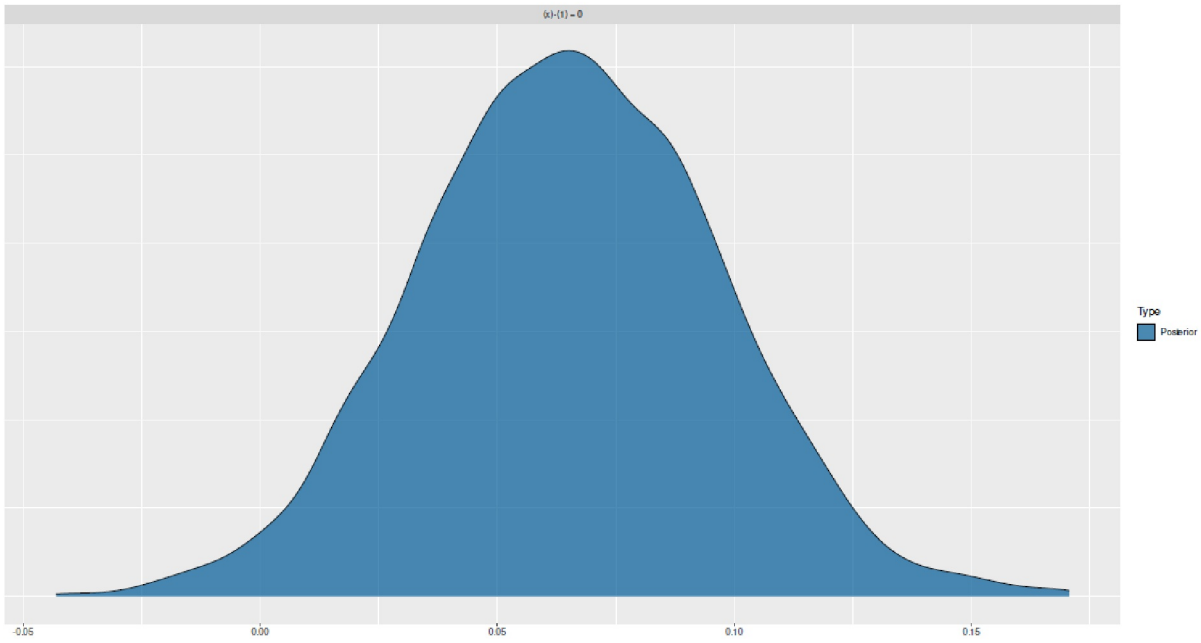


Figura 6.52. *Modelo lineal (ajuste casi perfecto, prueba de hipótesis b = 1)*

Ahora la regresión lineal se convierte en un diseño de tratamiento de grupo de control con dos grupos. Esto requiere un grupo de factores dicotómicos, que se crea de forma que ambos grupos difieran con respecto a la variable dependiente, pero no estén completamente separados. El grupo Control recibe $y_C \sim N(100, 10)$ y el grupo Tratamiento $y_T \sim N(105, 10)$. Las varianzas de los grupos no difieren, pero las medias sí. El tamaño del grupo *N* es N = 100 en cada caso. La figura 6.53 muestra los datos.

```
# case 2 - treatment design
set.seed(1172233)
ngroup <- 2
n <- 100
mu1 <- 100
mu2 <- 105
# same variances
sigma1 <- 10
sigma2 <- 10
yC <- rnorm(n,mu1,sigma1)
yT <- rnorm(n,mu2,sigma2)
y <- c(yC,yT)
group <- rep(c("Control","Treatment"),each=n)
xymodel <- data.frame(y=y,group=factor(group))
```

Ahora se repite los pasos anteriores y calcula la *d de Cohen*:

```
> # descriptive statistics
> do.call("rbind",with(xymodel, tapply(y, INDEX=group,
+ FUN=function(x) c(summary(x),SD=sd(x),VAR=var(x),fivenum2(x)))))
            Min.     1st Qu.    Median     Mean     3rd Qu.   Max.
Control     77.53147 93.27565  98.43336 100.0410 107.0461 125.3893
Treatment   84.27134 96.74215 104.96905 104.2391 109.5409 129.6605
           SD       VAR       minimum  lower-hinge median
Control   10.154850 103.12099 77.53147 93.22525     98.43336
Treatment 9.641469  92.95792 84.27134 96.64665    104.96905
```

```
           upper-hinge maximum
Control     107.0670   125.3893
Treatment   109.7786   129.6605
> cohensd(xymodel$y[xymodel$group == "Control"],
+ xymodel$y[xymodel$group == "Treatment"])
d|mean sd    d|pooled sd
0.4239915    0.4239915
```

y el modelo frecuentista:

```
# frequentist solution
lm.xymodel.CT <- lm(y ~ group, data=xymodel)
summary(lm.xymodel.CT)
display(lm.xymodel.CT)
par(mfrow=c(2,2))
plot(lm.xymodel.CT, col="purple")
```

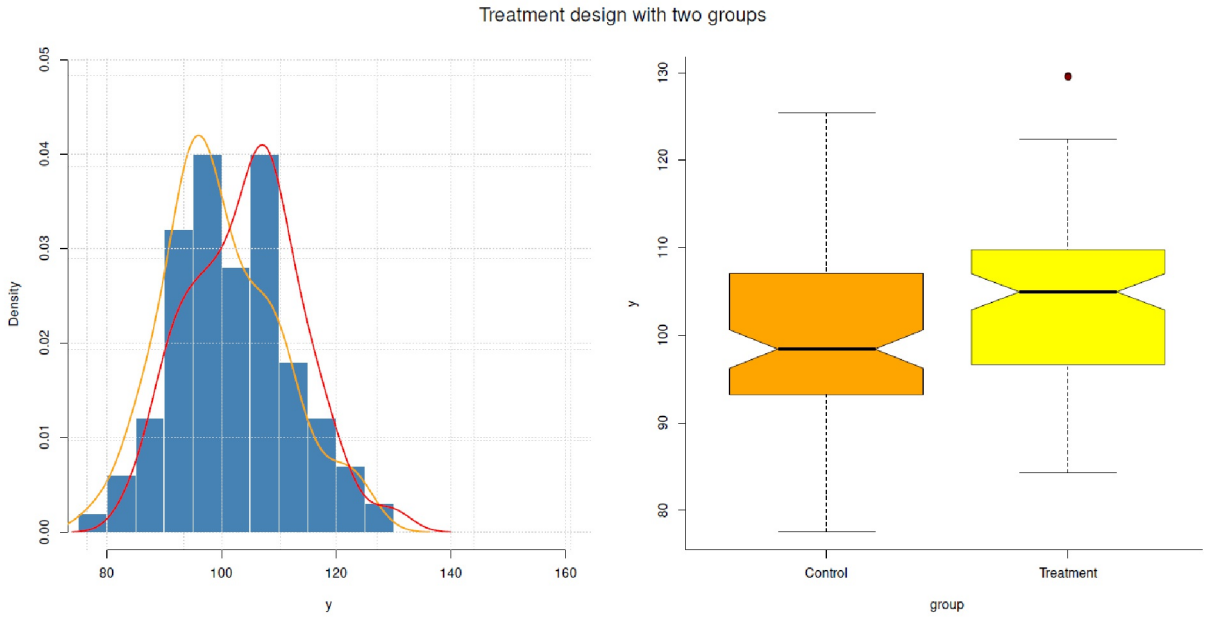


Figura 6.53. *Modelo lineal (dos grupos, histograma y boxplot)*

Es interesante comprobar la diferencia entre los dos grupos. El factor grupo es estadísticamente significativo según las reglas habituales de la convención, el ajuste del modelo con $R^2 = 0.039$ modesto, lo que se debe a la fuerte dicotomización de los datos mediante el factor grupo. El análisis bayesiano mediante `brm()` muestra con

```
Family: gaussian R-Output
Links: mu = identity; sigma = identity
Formula: y ~ group
Data: xymodel (Number of observations: 200)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
total post-warmup samples = 4000
Population-Level Effects:
              Estimate Est.Error l-95% CI u-95% CI Rhat
Intercept      100.15      0.98     98.28    102.10  1.00
groupTreatment   3.93      1.37      1.23      6.62  1.00
              Bulk_ESS Tail_ESS
Intercept         4022 3037
```

```
groupTreatment     3990 2740
Family Specific Parameters:
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma         9.93       0.49     9.02     10.96   1.00     4056 2918
Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS and Tail_ESS
are effective sample size measures, and Rhat is the potential scale reduction factor
on split chains (at convergence, Rhat = 1).
```

ninguna conspicuidad de las Posteriores. Las estimaciones del modelo corresponden aproximadamente al modelo frecuentista. Se puede ver en el gráfico de la distribución predictiva (véase la Fig. 6.54) con `pp_check()`, que la distribución empírica muestra acercamientos para dos picos y la distribución predictiva posterior un pico a causa del factor de grupo dicotómico.

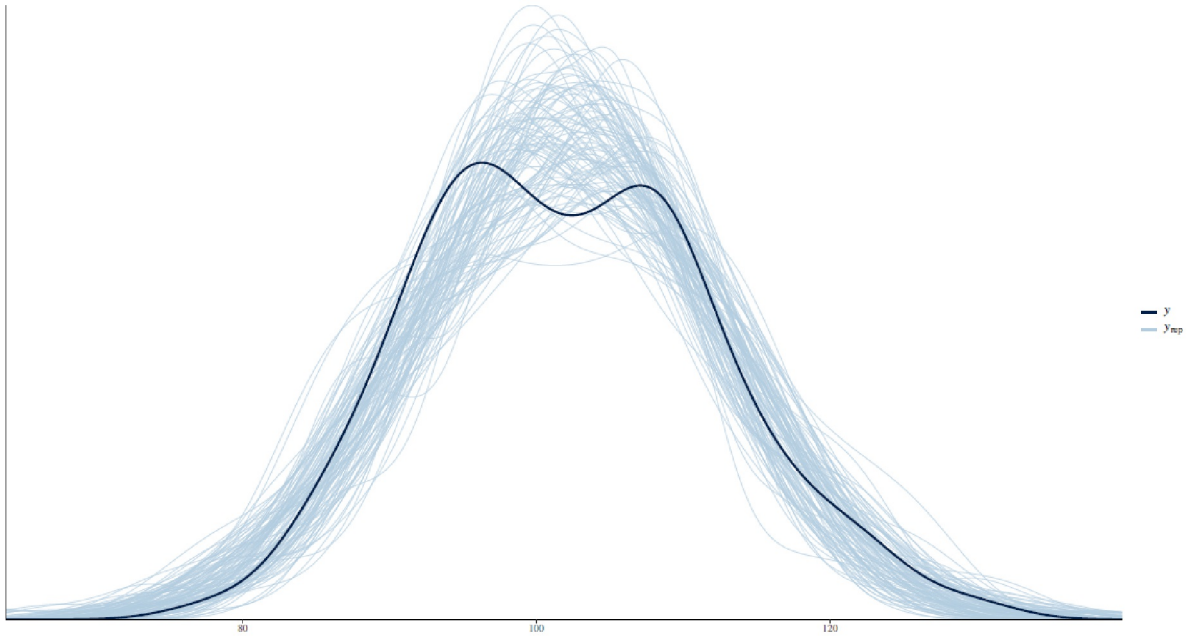


Figura 6.54. *Modelo lineal (dos grupos, distribución predictiva posterior)*

```
pp_check.brmsfit(brm.xymodel.CT, nsamples=100)
```

Anteriormente, se suponían las mismas varianzas por grupo. Ahora se cambia el conjunto de datos para que, además de los valores medios, las varianzas también sean heterogéneas, es decir, el grupo Control con y_C *~N(100, 14)* y el grupo Tratamiento con y_T *~N(110, 10)*. Esto puede hacerse de diferentes maneras, por lo que las diferencias deben observarse sobre todo en lo que respecta a las interpretaciones de los parámetros (véase la Fig. 6.55).

```
# case 3 - heterogenous variances between groups
set.seed(1172233)
ngroup <- 2
n <- 100
mu1 <- 100
sigma1 <- 14#12
mu2 <- 110
sigma2 <- 10
yC <- rnorm(n,mu1,sigma1)
yT <- rnorm(n,mu2,sigma1)
yT.alt <- rnorm(n,mu2,sigma2)
# different mu, same variance
```

```
y <- c(yC,yT)
# different mu, different variance
y.alt <- c(yC,yT.alt)
group <- rep(c("Control","Treatment"),each=n)
xymodel <- data.frame(y=y,y.alt=y.alt,group=factor(group))
```

`bf()` es una función de ayuda de R de `brms` que maneja estas formulaciones de modelos. La llamada `get_prior()` alrededor del exterior muestra qué Priors se van a asignar. Este es un buen comienzo para entender el contenido del modelo. Los detalles se pueden encontrar en las viñetas y tutoriales del paquete `brms` de R. El siguiente primer modelo no tiene en cuenta las diferentes varianzas:

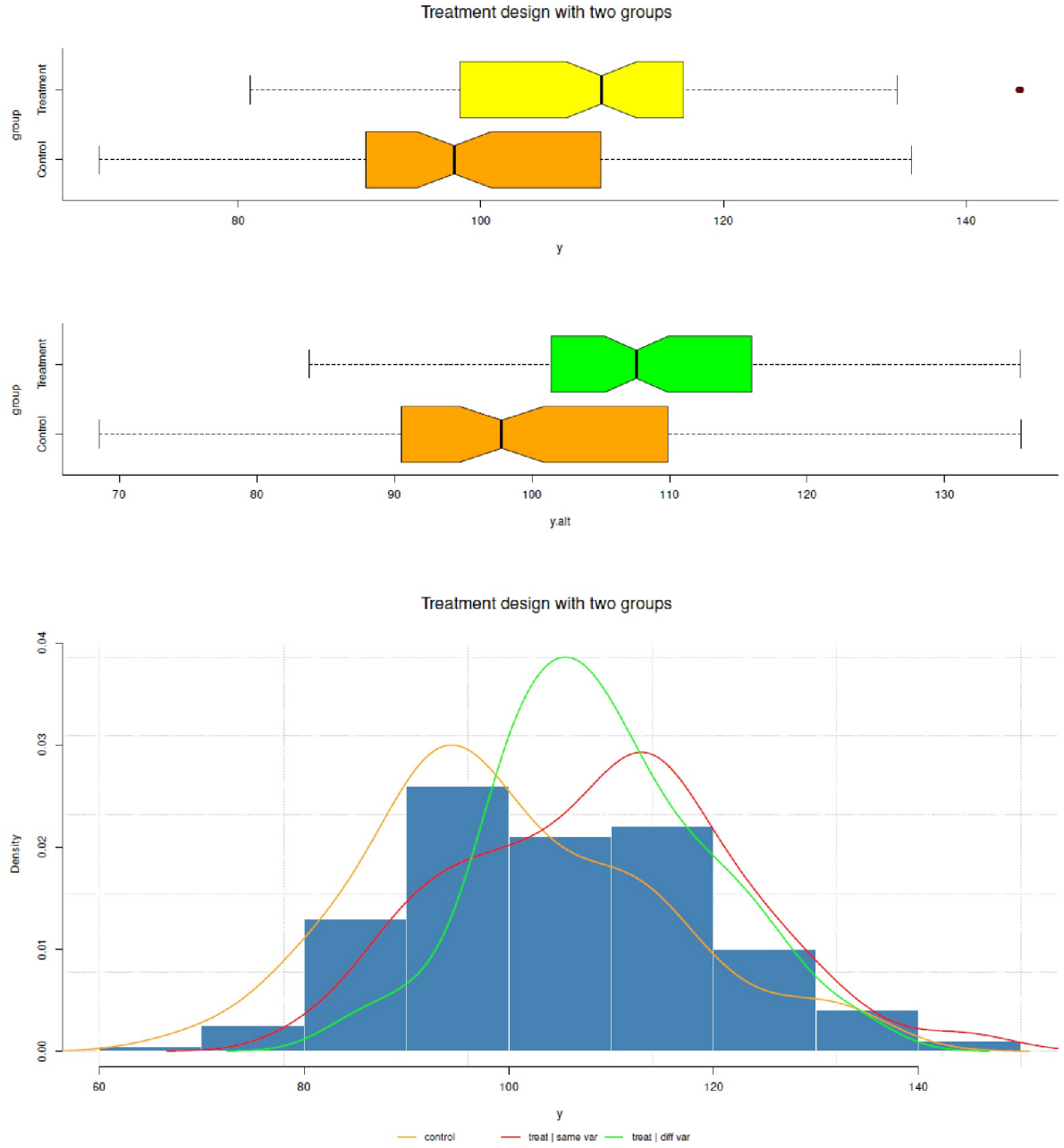


Figura 6.55. *Modelo lineal (dos grupos, varianzas iguales y diferentes, boxplot e histograma)*

```
# different mu, different variance
# non-heterogenous variances between groups
get_prior(bf(y.alt ~ group), data=xymodel)
# define priors for non-heterogenous variances
P.nonhet <- c(prior(normal(1,5), class="b", coef="groupTreatment"),
       prior(student_t(2,0,10), class="Intercept"),
       prior(student_t(2,0,10), class="sigma"))
```

Pero este modelo sí lo hace:

```
# different mu, different variance
# heterogenous variances between groups
# one prior for the variances
# different priors
# group factor tests also SD of residual of the response
get_prior(bf(y.alt ~ group, sigma ~ group), data=xymodel)
```

El modelo así formulado indica que el factor grupo examina tanto las diferencias de medias como las de las varianzas residuales de las variables dependientes, por separado unas de otras. Se podrían formular varianzas separadas para los dos grupos:

```
# model different sigmas for each group
get_prior(bf(y.alt ~ group, sigma ~ 0 + group), data=xymodel)
```

La comparación de las Priores en la salida, que al mismo tiempo incluyen todos los parámetros del modelo a estimar, muestra que en el primer caso el `groupTreatment` sólo aparece como predictor. En el segundo modelo aparece el `groupTreatment` como factor de diferencias en `sigma`, y en el último modelo se estiman tanto el `groupControl` como el `groupTreatment` por separado `sigma`. El término `+ 0` elimina el `intercepto` normal para `sigma`. Si se quiere modelar esto directamente con la propia Prior, se escribe con el término `intercepto` reservado:

```
# model the intercept and allow to specify a prior on it
# the '0 +' takes out the default intercept
# (= like "-1' in other R models)
# the reserved term 'intercept' shows you mean
# the regular/ real intercept
# ie. you model the intercept as a unique factor in the model
# ?brmsformula
get_prior(bf(y.alt ~ group, sigma ~ 0 + Intercept + group),
       data=xymodel)
# + 0 = make one group as the intercept, ie. shift model
# to this group etc. as a base
# define priors for heterogenous variances
P.hetv <- c(prior(normal(1,5), class="b", coef="groupTreatment"),
       prior(student_t(2,0,10), class="Intercept"),
# one prior on sigma
prior(student_t(2,0,10), class="b", dpar="sigma",
       coef="groupTreatment"),
prior(student_t(3,0,10), class="Intercept", dpar="sigma")
)
```

Ahora vemos que `groupControl` ha desaparecido de nuevo para `sigma`, pero ahora el `intercept` aparece como su propio factor con `class=b` (= beta) y su propia Prior. No entramos en todas las posibilidades de análisis de estos diferentes modelos y sus interpretaciones, así como en las variantes de asignación de Priors. Por defecto, `brm()` utiliza Priors difusas. En lo que sigue, nos limitamos a estimar los modelos con `family=student` en lugar de `family=gaussian`, es decir, utilizamos una distribución *t* más robusta. En principio estamos tratando con cuatro modelos:

```
# robust estimation with family="student" R-Code
# use gaussian() for other family
# no handling of variances
brm.xymodel.CT.nonhetv <- brm(bf(y.alt ~ group), data=xymodel,
        family="student",
        sample_prior=TRUE,
        save_all_pars=TRUE, prior=P.nonhet)
# models the effect of group on the mean AND
# the residual standard deviation of the response distribution
# ie. it is like a group-test how the group influences
# the SD of the residuals on the response
# it is not modelling the variances differently,
# ie. different variances for each group
# the latter uses 'sigma ~ 0 + group'
# group factor for sigma
brm.xymodel.CT.hetv <- brm(bf(y.alt ~ group, sigma ~ group),
        data=xymodel, family="student",
        sample_prior=TRUE, save_all_pars=TRUE,
        prior=P.hetv)
# separate variances for groups
brm.xymodel.CT.hetv1 <- brm(bf(y.alt ~ group, sigma ~ 0 + group),
        data=xymodel, family="student",
        sample_prior=TRUE, save_all_pars=TRUE)
# unique estimation of intercept with prior
brm.xymodel.CT.hetv2 <- brm(bf(y.alt ~ group,
        sigma ~ 0 + intercept + group),
        data=xymodel, family="student",
        sample_prior=TRUE, save_all_pars=TRUE)
```

En general, se observa en los modelos que las probabilidades posteriores de los parámetros son relativamente simétricas y de un solo pico. He aquí el modelo con varianzas heterogéneas:

```
> # all effects
> fixef(brm.xymodel.CT.hetv)
                     Estimate    Est.Error   Q2.5      Q97.5
Intercept            100.3765004 1.39540302 97.6121991 103.15263076
sigma_Intercept        2.6167763   0.07951924  2.4598039   2.77178100
groupTreatment         8.1430637   1.71717500  4.7446335  11.59638220
sigma_groupTreatment  -0.2976572   0.10582405 -0.5047689 -0.09212344
> # normal effects
> fixef(brm.xymodel.CT.hetv)[c(1,3),]
                  Estimate Est.Error Q2.5       Q97.5
Intercept        100.376500 1.395403 97.612199 103.15263
groupTreatment     8.143064 1.717175  4.744633  11.59638
> # sigmas have to be exp()
> # intercept is the sigma (standard deviation) of group 0
> # groupTreatment is the effect of group 1 on sigma (of group 0)
> # standard deviation of groupTreatment is
> # sigma_Intercept + sigma_groupTreatment
> # first we check the direction of the effect
> sigmaef <- fixef(brm.xymodel.CT.hetv)[c(2,4),]
> sigmaef
                      Estimate  Est.Error   Q2.5       Q97.5
sigma_Intercept       2.6167763 0.07951924  2.4598039  2.77178100
sigma_groupTreatment -0.2976572 0.10582405 -0.5047689 -0.09212344
> # now exp()
> sigmaef.exp <- exp(sigmaef[,"Estimate"])
> sigmaef.exp
sigma_Intercept sigma_groupTreatment
13.6915148      0.7425558
> # group Control
> sigmaef.exp[1]
sigma_Intercept
```

```
13.69151
> # group Treatment
> # see above - negative effect of groupTreatment on SD
> ifelse(sigmaef[2,"Estimate"] > 0,
+ sum(sigmaef.exp), sigmaef.exp[1]-sigmaef.exp[2])
[1] 12.94896
> sigmaef.exp[1] + sign(sigmaef[2,1])*sigmaef.exp[2]
sigma_Intercept
12.94896
> # compare to SDs original data
> with(xymodel, tapply(y.alt, group, sd))
Control  Treatment
14.21679 10.58239
```

Además de los efectos fijos, obtenemos las varianzas siguiendo las líneas anteriores con el fin de obtenerlas para los grupos de control y tratamiento respectivamente. La primera pregunta que nos hacemos es si vale la pena considerar las varianzas. La distribución predictiva posterior representa la adecuación del modelo (véase la Fig. 6.56):

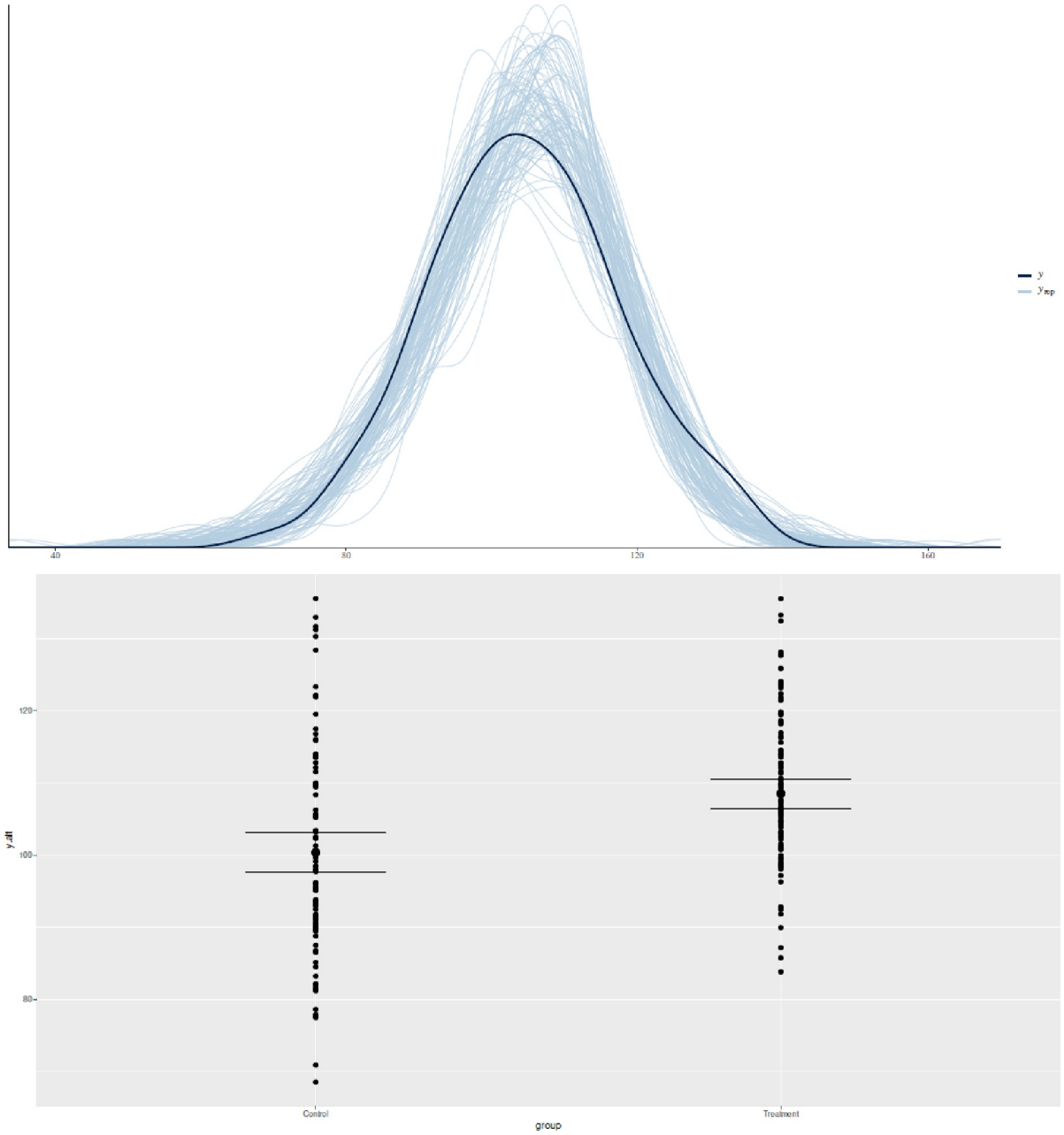


Figura 6.56. *Modelo lineal*
(dos grupos, varianzas diferentes, distribución predictiva posterior y efectos marginales)

```
# posterior predictive checks
pp_check(brm.xymodel.CT.nonhetv, nsamples=100)
```

Para ello, se trazan los efectos de los grupos por separado (véase la Fig. 6.56 arriba)

```
plot(marginal_effects(brm.xymodel.CT.hetv), points=TRUE)
```

y se nota, que es necesario separar las varianzas, ya que obviamente difieren entre sí en su magnitud. Ahora se formula una hipótesis que examina la influencia del factor grupo sobre `sigma`.

```
# test hypothesis about different variances
# Evid.Ratio is a ratio (BF_01)
> h1.hetv <- hypothesis(brm.xymodel.CT.hetv,
        c("sigma_Intercept = 0", "sigma_Intercept +
        sigma_groupTreatment = 0"))
> h1.hetv
Hypothesis Tests for class b:
                     Hypothesis Estimate Est.Error CI.Lower CI.Upper
1 (sigma_Intercept) =      0       2.62       0.08    2.46     2.77
2 (sigma_Intercept+... =   0       2.32       0.08    2.16     2.48
                     Hypothesis Evid.Ratio Post.Prob Star
1 (sigma_Intercept) =      0       NA         NA *
2 (sigma_Intercept+... =   0       NA         NA *
---
'CI': 90%-CI for one-sided and 95%-CI for two-sided hypotheses.
'*': For one-sided hypotheses, the posterior probability exceeds 95%;
for two-sided hypotheses, the value tested against lies outside
the 95%-CI. Posterior probabilities of point hypotheses assume
equal prior probabilities.
> plot(h1.hetv, ignore_prior=TRUE)
```

Esta hipótesis `h1.hetv` examina por separado si la sigma del `intercepto` o la `sigma` del `intercepto` juntos con `groupTreatment` es cero o, en otras palabras, si las varianzas del grupo de control o de tratamiento difieren de cero. Esto no es así en cada caso, como se desprende del factor de Bayes, que puede leerse en la columna `Evid.Ratio` – es decir, ambas hipótesis asumen valores no iguales a cero, con la máxima probabilidad, si se calcula 1/BF, que aquí corresponde a $1/0 = \infty$. Además, obtenemos los límites de confianza del 95%.

La hipótesis `h1` puede combinarse en la hipótesis `h2`. Esta `h2` investiga si `sigma` del grupo de tratamiento es menor que el nivel del grupo de control.

```
> # test that the treatment group variance is larger
> # than the control group variance
> hypo2 <- c("exp(sigma_Intercept +
        + sigma_groupTreatment) < exp(sigma_Intercept)")
> h2.hetv <- hypothesis(brm.xymodel.CT.hetv, hypo2)
> h2.hetv
Hypothesis Tests for class b:
                     Hypothesis Estimate Est.Error CI.Lower CI.Upper
1 (exp(sigma_Interc... < 0        -3.54      1.29    -5.72    -1.47
                     Hypothesis Evid.Ratio Post.Prob Star
1 (exp(sigma_Interc... < 0        306.69     1         *
---
'CI': 90%-CI for one-sided and 95%-CI for two-sided hypotheses.
'*': For one-sided hypotheses, the posterior probability exceeds 95%;
for two-sided hypotheses, the value tested against lies outside
the 95%-CI. Posterior probabilities of point hypotheses assume
equal prior probabilities.
> BF01 <- h2.hetv$hypothesis$Evid.Ratio
> BF01
[1] 306.6923
> # BF10
> 1/BF01
[1] 0.003260597
> plot(h2.hetv, ignore_prior=TRUE)
```

Este también es el caso, y no es sorprendente, ya que el grupo de control se generó con una media menor pero varianza significativamente mayor que la del grupo de tratamiento. En consecuencia, la prueba unilateral

es correcta. Si se utiliza el modelo `brm.xymodel.CT.hetv1`, las varianzas de los dos grupos pueden representarse una al lado de la otra (véase la Fig. 6.57).

```
# then you can plot both sigmas of the treatments against each other
# plot sigmas
str(posterior_samples(brm.xymodel.CT.hetv1))
sigmas <- exp(posterior_samples(brm.xymodel.CT.hetv1, "^b_sigma_"))
ggplot(stack(sigmas), aes(values)) + geom_density(aes(fill = ind))
```

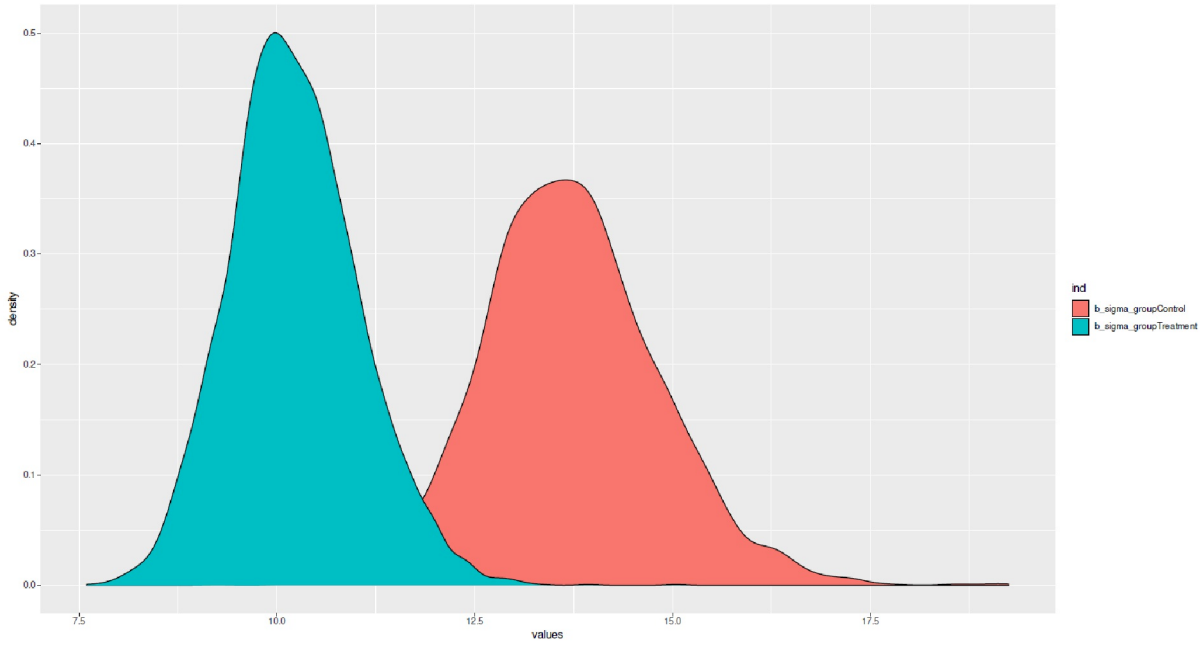


Figura 6.57. *Modelo lineal – dos grupos, varianzas diferentes (varianzas posteriores)*

Los valores que deben leerse corresponden a la forma en que se generaron los datos (véase más arriba). Se podría formular una hipótesis dirigida `hipo3` si la varianza del grupo de tratamiento es menor que la del grupo de control.

```
> hypo3 <- c("exp(sigma_groupTreatment) < exp(sigma_groupControl)")
> hypo3.hetv <- hypothesis(brm.xymodel.CT.hetv1, hypo3)
> hypo3.hetv
Hypothesis Tests for class b:
                    Hypothesis Estimate Est.Error CI.Lower CI.Upper
1 (exp(sigma_groupT... < 0        -3.49      1.34    -5.73    -1.38
                    Hypothesis Evid.Ratio Post.Prob Star
1 (exp(sigma_groupT... < 0        332.33     1          *
---
'CI': 90%-CI for one-sided and 95%-CI for two-sided hypotheses.
'*': For one-sided hypotheses, the posterior probability exceeds 95%;
for two-sided hypotheses, the value tested against lies outside
the 95%-CI. Posterior probabilities of point hypotheses assume
equal prior probabilities.
```

Obviamente, éste es el caso. La comprobación gráfica predictiva posterior mediante

```
pp_check(brm.xymodel.CT.hetv1, nsamples=100, type="error_scatter_avg")
```

muestra que el modelo explica suficientemente bien los datos. También es interesante comparar los modelos anteriores. Para ello, se pueden calcular distintos criterios de calidad (véase también el capítulo 6.8.2 sobre los criterios de información). En primer lugar, hay que hacerlo por separado para cada modelo

```
brm.xymodel.CT.hetv.ic1 <- add_criterion(brm.xymodel.CT.hetv1,
        criterion=c("loo","waic","kfold","R2","marglik"))
```

y después se puede comparar los modelos:

```
# compare models
for(i in c("loo","waic","kfold"))
{
cat("\n",i)
print(loo_compare.brmsfit(brm.xymodel.CT.nonhetv.ic,
        brm.xymodel.CT.hetv.ic,
        brm.xymodel.CT.hetv.ic1, criterion=i))
}
```

También se puede mostrar el ajuste de calidad R^2 incluyendo cuantiles con

```
# R2
bayes_R2.brmsfit(brm.xymodel.CT.nonhetv)
bayes_R2.brmsfit(brm.xymodel.CT.hetv)
bayes_R2.brmsfit(brm.xymodel.CT.hetv1)
bayes_R2.brmsfit(brm.xymodel.CT.hetv2)
```

Estos difieren mínimamente entre sí y se sitúan aproximadamente en el intervalo de 0.098 a 0.119 con una tolerancia de error de 0.039 a 0.04. Se obtiene una visión general de las muestras posteriores para parámetros específicos con

```
samps0 <- posterior_samples(brm.xymodel.CT.hetv)
```

y puede seguir procesándolas. Las hipótesis pueden comprobarse, trazarse, etc. (véase la Fig. 6.58):

```
str(samps0) R-Code
samps0.d <- dim(samps0)
d12 <- ceiling(sqrt(samps0.d[2]))
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l", mfrow=c(d12,d12))
fac <- 1.15
for(i in 1:samps0.d[2])
{
print(i)
dens <- density(samps0[,i])
ylim <- c(0,max(dens$y)*fac)
hist(samps0[,i], prob=TRUE, pre.plot=grid(), bg="grey",
        ylim=ylim, xlab=colnames(samps0)[i],main="")
lines(dens, col="darkred", lwd=2)
}
mtext("Samples of posterior of model parameters",
        outer=TRUE, line=-2, cex=1.5, side=3)
```

Hasta aquí el análisis de modelos lineales mediante la estadística de Bayes. Por lo demás, el paquete `brms` de R permite un número increíble de otras posibilidades. Como de costumbre, se puede extraer información de los Posteriors o de los modelos, lo que también es habitual en otros análisis de modelos en R. Se obtiene aún más si se busca directamente en los objetos generados por el análisis:

```
# extract general information for fitted values
fitted(brm.xymodel.CT.hetv)
# fixed effects
fixef(brm.xymodel.CT.hetv)
# random effects - here not present
ranef()
# log likelihood
log_lik(brm.xymodel.CT.hetv)
# log posterior
log_posterior(brm.xymodel.CT.hetv)
# posterior model probabilities from marginal likelihoods
post_prob(brm.xymodel.CT.hetv, brm.xymodel.CT.nonhetv)
# posterior intervals Q2.5 and Q97.5
posterior_interval(brm.xymodel.CT.hetv)
# posterior table per model paramter
posterior_table(brm.xymodel.CT.hetv)
# residuals
residuals(brm.xymodel.CT.hetv)
# variance-covariance matrix
vcov(brm.xymodel.CT.hetv)
```

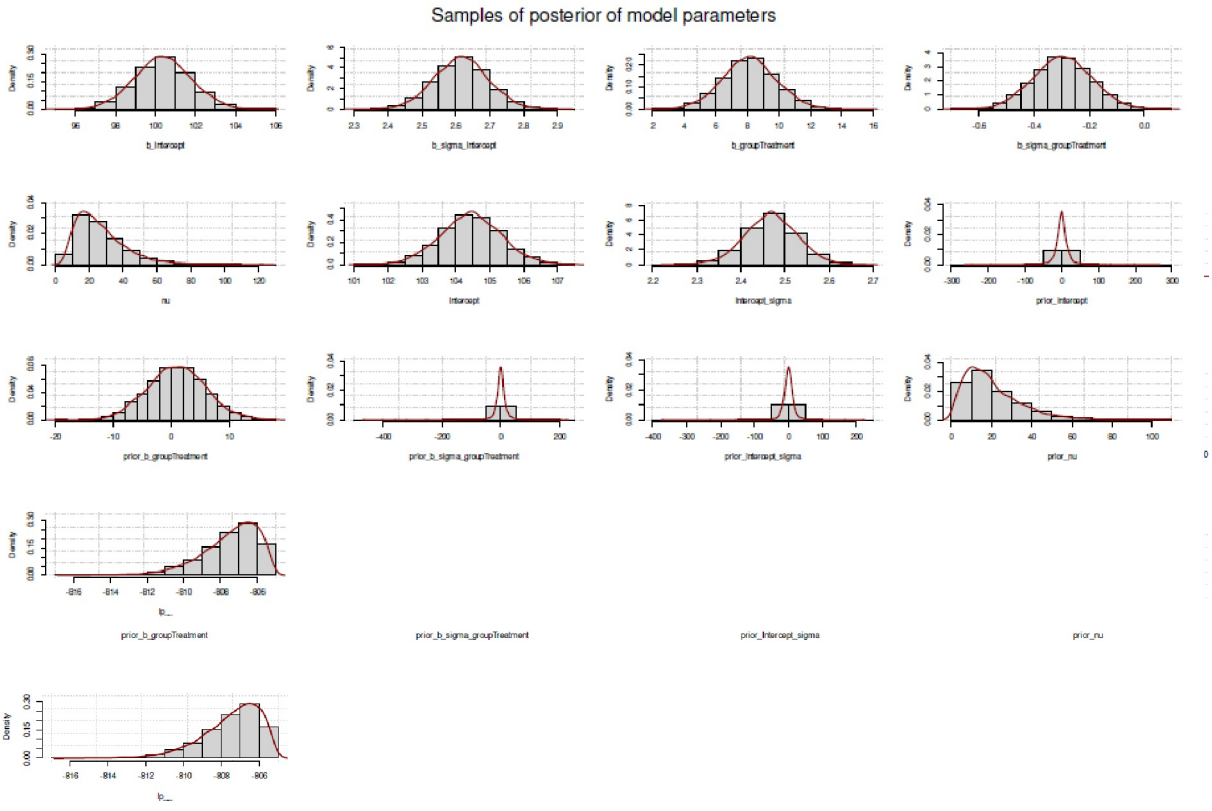


Figura 6.58. *Modelo lineal (dos grupos, diferentes varianzas, Posteriors)*

A continuación se presenta un estudio de caso de comprobaciones predictivas posteriores, basado en el trabajo de Kruschke (2015b) y JAGS.

6.8.4.6 Estudio de caso – humor y la producción de palabras

En el capítulo 6.8.1 ya se mencionó el estudio de Gürtler (2005). Sirve como otro ejemplo de comprobaciones predictivas posteriores, específicamente para el caso continuo desde dentro de R. En ese caso, utilizamos los scripts de R `DBDA2E-utilities.R` y `Jags-Ymet-XmetMulti-Mrobust.R` (Kruschke, 2016d) del libro de Kruschke (2015b), que arreglan la llamada a JAGS. La forma más sencilla de crear el modelo es utilizar la conocida notación de R y transformarlo en un modelo basado en la codificación ficticia mediante `model.matrix()` (Fox, 2002). En primer lugar, sin embargo, examinamos descriptivamente los propios datos (véase la Fig. 6.59) y observamos el código R correspondiente (`ptII_quan_Bayes_case_word-counts-PPC.r`).

```
# get data
diss <- read.csv("LG_school-words-raw.tab", header=TRUE, sep="\t")
str(diss)
head(diss)
tail(diss)
namen <- names(diss)
with(diss, cor(W.noSC,uW.noSC))
with(diss, cor(W.wSC,uW.wSC))
with(diss, xtabs(age ~ schooltype +sex))
par(mfrow=c(2,3))
with(diss, plot(W.noSC,uW.noSC, bty="n", col="violetred3",
       lty=1, lwd=1, type="p", pre.plot=grid()))
with(diss, lines(lowess(uW.noSC ~ W.noSC), col="blue", lwd=2, lty=2))
with(diss, plot(W.noSC,age, bty="n", col="violetred3",
       lty=1, lwd=1, type="p", pre.plot=grid()))
with(diss, boxplot(W.noSC ~ sex * schooltype, notch=TRUE, col="violetred3"))
with(diss, interaction.plot(schooltype, sex, W.noSC,
       bty="n", pre.plot=grid(), col=c("darkred","blue")))
with(diss, interaction.plot(sex, schooltype, W.noSC,
       bty="n", pre.plot=grid(), col=c("darkred","blue")))
```

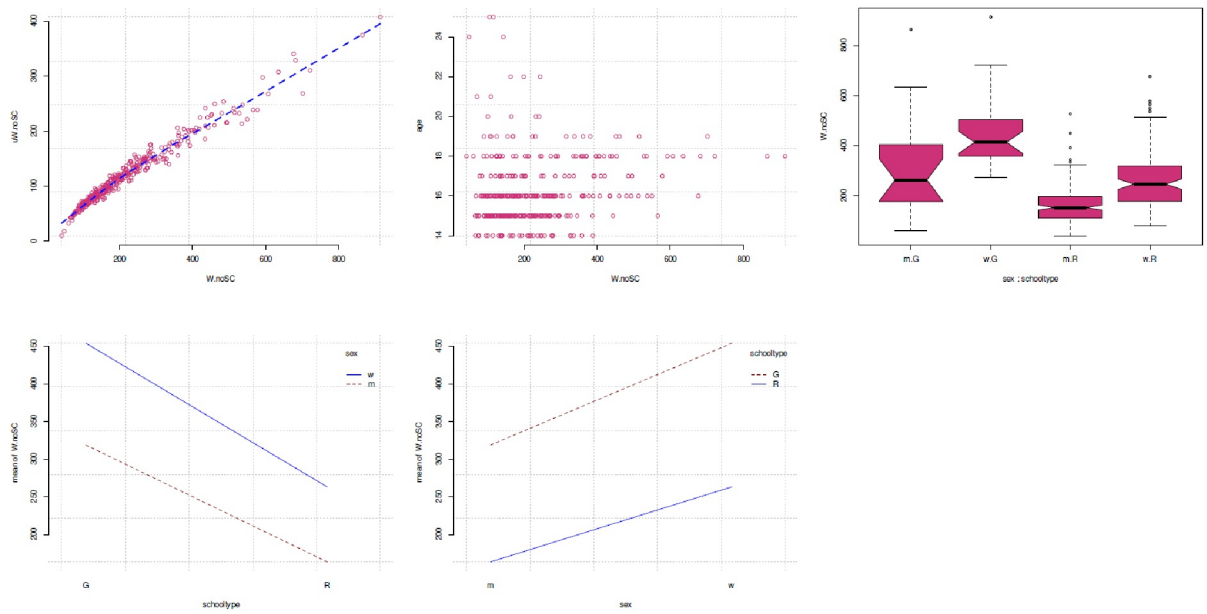


Figura 6.59. *Estudio de Gürtler (2005, humor – producción de palabras humor, gráficos exploratorias parte 1).*

De ello se deduce que las chicas escriben más que los chicos y que se escribe más en el Gymnasium que en la Realschule. Todo esto no dice nada sobre lo que se escribe y con qué calidad e intención. El diagrama de dispersión proporciona más información. Para evitar los datos de edad artificialmente discretos – nadie tiene exactamente 15, 16 o 17 años, sino 16 años y tantos días – utilizamos `jitter()`, que desplaza ligeramente los valores. El procedimiento ya se ha descrito y aplicado en el capítulo 6.8.1. Obviamente, la producción de palabras crece con la edad, luego hace una brusca curva descendente en torno a los 19 años, y después se estabiliza en un nivel bajo. Una mirada a la tabulación de la edad, el sexo y el tipo de escuela ya explica esto parcialmente.

```
> with(diss, ftable(sex,stype,age))
        age 14 15 16 17 18 19 20 21 22 24 25
sex stype
m   B        0  0  2  9  8  8  4  2  2  2  2
    R        0  0  0  2 13  3  0  0  0  0  0
    G       18 75 54  9  5  2  0  0  0  0  0
w   B        0  0  1  1  0  0  0  0  1  0  0
    R        0  0  0  5 19  4  0  0  0  0  0
    G       16 46 40  4  3  0  0  0  0  0  0
```

Alternativamente, se obtiene información comparable con otros gráficos de interacción, que siempre son útiles para las comparaciones de grupos (véase la Fig. 6.60).

```
# prepare data
diss$W.noSC <- as.integer(diss$W.noSC)
diss$W.noSC.log <- log(diss$W.noSC)
diss$age.log <- log(diss$age)
diss$SS <- factor(diss$SS)
diss$sex <- factor(diss$sex)
diss$schooltype <- factor(diss$schooltype)
# as.numeric(factor(paste(diss$sex,diss$schooltype,sep="")))
diss$SSn <- as.numeric(factor(diss$SS))
diss$sexn <- as.numeric(factor(diss$sex))
diss$stypen <- as.numeric(factor(diss$schooltype))
diss$age.cat <- cut(as.numeric(diss$age), breaks=c(13,16,19,25),
       include.lowest=TRUE, right=TRUE,
       labels=c("(14-16]","(17-19]","(20-25]"))
diss$age.cat1 <- cut(as.numeric(diss$age), breaks=c(13,19,25),
       include.lowest=TRUE, right=TRUE,
       labels=c("(14-19]","(20-25]"))
str(diss)
# interaction plots
colos <- colorRampPalette(c("red","green","orange","blue"))
with(diss, interaction.plot(stype, sex, W.noSC, bty="n",
       pre.plot=grid(), col=colos(2)))
with(diss, interaction.plot(sex, stype, W.noSC, bty="n",
       pre.plot=grid(), col=colos(3)))
with(diss, interaction.plot(stype, sex, W.noSC.log, bty="n",
       pre.plot=grid(), col=colos(2)))
with(diss, interaction.plot(sex, stype, W.noSC.log, bty="n",
pre.plot=grid(), col=colos(3)))
with(diss, interaction.plot(stype, sex, W.noSC.log, bty="n",
       pre.plot=grid(), col=colos(3)))
with(diss, interaction.plot(age.cat, stype, W.noSC.log, bty="n",
       pre.plot=grid(), col=colos(3)))
with(diss, interaction.plot(age.cat, sex, W.noSC.log, bty="n",
       pre.plot=grid(), col=colos(2)))
```

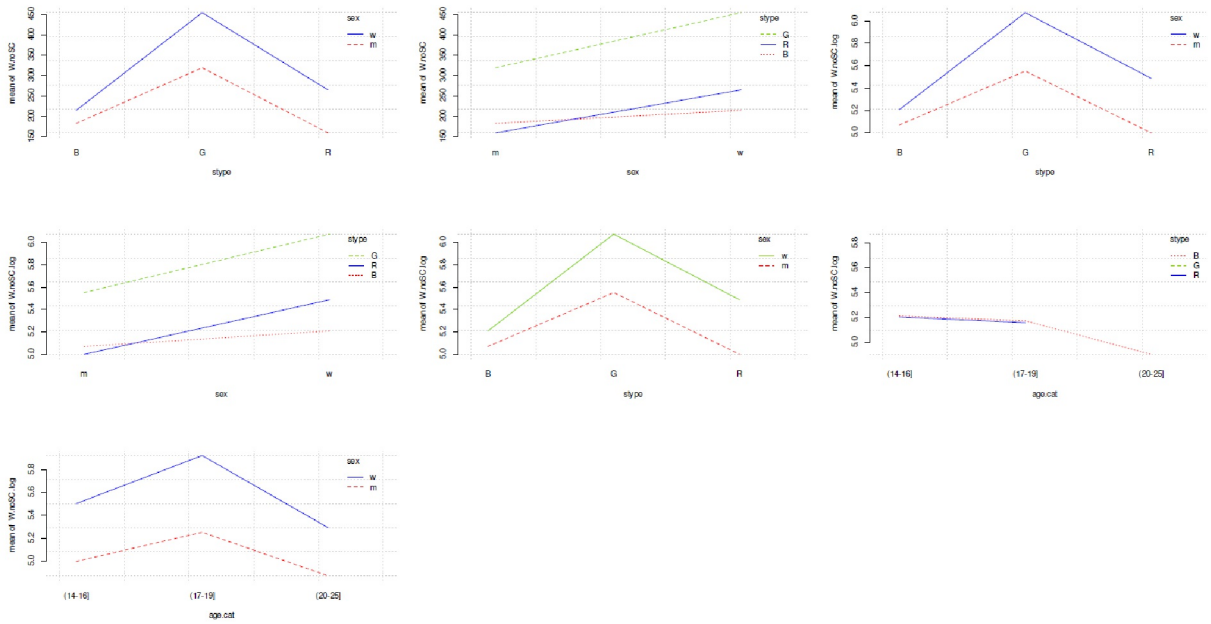


Figura 6.60. *Estudio de Gürtler*
(2005, humor – producción de palabras humor, gráficos exploratorias parte 2).

Debe prestarse especial atención a la muestra: se observa que los estudiantes del Gymnasium en la franja de edad de 14-16 años están completamente ausentes y (condicionalmente) a los 17 años siguen estando claramente ausentes. Por el contrario, la Realschule termina después del 10º curso, por lo que los jóvenes de más edad están poco o nada representados aquí. La variabilidad es, por tanto, limitada en ambos tipos de escuela. En el Gymnasium se podría cubrir la franja de edad más joven, pero no en esta muestra (selectiva).

En cambio, la Escuela profesional con el objetivo de obtener un certificado de Realschule muestra una variabilidad de edad más amplia, concretamente de 16 a 25 años e incluye sobre todo a jóvenes y (jóvenes) varones. En el Gymnasium (con un tamaño de muestra similar en términos absolutos al de la Escuela profesional) y en la Realschule (muestra significativamente mayor que la de la Escuela profesional y el Gymnasium juntos) hay más chicos que chicas. Por tanto, sospechamos que una interacción entre el sexo y el tipo de escuela, así como una distribución truncada en todo el espectro de edad. Así que antes de entrar en las correlaciones con la producción de palabras, necesitamos esta información para comprender mejor el diagrama de dispersión de la edad y la producción de palabras (véase la Fig. 6.59). Merece la pena considerar aquí si no se trata de un modelo compuesto. Sin embargo, esto distraería mucho del problema básico de las comprobaciones predictivas posteriores. Una vez más, este estudio muestra la necesidad de *examinar con precisión una muestra antes de pasar al contenido*. En sentido estricto, aquí faltan áreas de muestra necesarias, es decir, la muestra está incompleta y no es representativa. Por un lado, los conjuntos de datos sobre los tipos de escuelas están incompletos y, por tanto, la representación de género del mismo modo. En consecuencia, existen lagunas en el espectro de edades. Como era de esperar por sentido común la producción de palabras *no* aumenta infinitamente con la edad, pero sin duda muestra una estrecha correlación con el nivel de escolarización, la motivación y una disminución con el aumento de la edad, lo que se podría entender como una especie de saturación natural. Si fuera empíricamente posible recoger más datos, el primer objetivo sería ajustar el diseño con respecto al género, los grupos de edad y el tipo de escuela. Lo que quedaría sería la natural terminación de las distribuciones en la parte superior e inferior (véase también la composición de la muestra en el estudio de Wipfler, apartado 5.5.6).

Esto tiene varias consecuencias para la modelización. Una sería la transformación de escala, la cuestión de si un `log()` sobre la producción de palabras y la edad sería útil, es decir, si en la escala logarítmica la

edad y la producción de palabras muestran una correlación aproximadamente lineal. Además de la edad como covariable métrica, se necesitan también categorías de edad adicionales derivadas de ella, como 14-16, 17-19 y 20-25 años, para captar las distinciones más gruesas a nivel discreto. Esto sigue una idea de Dalgaard (2004). Esto puede comprobarse con factores anidados y factores bayesianos (Kruschke, 2013c). El modelado más detallado no se describe más aquí. Por curiosidad examinamos todas de las posibilidades enumeradas.

Ahora pasamos a la estimación del modelo con JAGS. El script de Kruschke necesita cierta información sobre las cadenas MCMC. Estos incluyen el número de pasos almacenados, la cantidad de adelgazamiento (para evitar la autocorrelación), realizado como un intervalo de adelgazamiento, los nombres de los predictores y de la variable dependiente, y la creación de la matriz del modelo con la función `model.matrix()` conocido de R. Esto ahorra la molestia de escribir el código JAGS directamente, lo que tendría ventajas, sin embargo, ya que uno puede entonces denotar directamente las llamadas cantidades de interés directamente en el código JAGS, por ejemplo para realizar comprobaciones predictivas posteriores directamente con JAGS (`ptII_quan_Bayes_case_wordcounts-PPC.r`).

```
# posterior predictive check with R after JAGS
# constants
fileNameRoot <- "Diss-"
numSavedSteps <- 1e4
thinSteps <- 50
numSavedSteps <- 1e3
nChains <- 3
# simple model
# response
yName1 <- "W.noSC.log"
# predictors
xName1 <- c("age.log","sexn","stypen")
diss1 <- diss[,c(yName1,xName1)]
naid <- which(is.na(diss1), arr.ind=TRUE)[,1]
diss.nona <- diss1[-naid,]
```

Un vistazo rápido a los datos brutos antes del análisis:

```
> head(diss.nona)
  W.noSC.log age.log      sexn stypen
1  5.598422   2.772589    1       2
2  5.513429   2.708050    2       2
3  4.919981   2.639057    1       2
4  5.501258   2.708050    2       2
5  5.446737   2.708050    1       2
6  5.602119   2.708050    2       2
```

Esto es seguido por la llamada con `genMCMC()`, que llama a JAGS y genera las cadenas MCMC, y por lo tanto estima el modelo.

```
# run model
mcmc.dm1 <- genMCMC( data=diss.nona, xName=xName1, yName=yName1,
        numSavedSteps=numSavedSteps, thinSteps=thinSteps,
        saveName=fileNameRoot, nChains=nChains)
str(mcmc.dm1)
```

El modelo resultante puede analizarse numérica y visualmente como de costumbre – aquí con el paquete `bayesplot` de R. El procedimiento se describe en Gabry (2018). Numéricamente, utilizamos una versión ligeramente modificada de `smryMCMC()` de Kruschke (2016d), renombrada como `smryMCMC2()`, que produce además las desviaciones estándar y las varianzas de la Posterior.

```
# analysis
# warning... creates a lot of plots
mcmc.diag.kruschke(model=mcmc.dm1, dats=diss.nona, xName=xName1, yName=yName1)
```

El parámetro ν corresponde a los grados de libertad, de modo que para $\nu \to \infty$ la distribución Student-*t* cambia a la distribución normal. Un ejemplo interesante es un gráfico de coordenadas en paralelo según Gabry, Simpson, Vehtari, Betancourt y Gelman (2019) con `mcmc_parcoord()` del paquete `bayesplot` de R, que muestra la distribución de los parámetros (véase la Fig. 6.61, con el parámetro ν, escalado abajo). Este gráfico muestra las extracciones MCMC por parámetro unidimensionalmente en paralelo. Éstos se pueden normalizar de antemano (véase la Fig. 6.62, escalada abajo), si las escalas de los parámetros son muy diferentes. En el gráfico, una única extracción de la cadena MCMC se extiende horizontalmente – como una línea conectada – a través de los parámetros representados. Antes creamos un modelo más complejo

```
# more complex model R-Code
diss.model <- cbind(W.noSC.log=diss.nona$W.noSC.log,
        with(diss, model.matrix(W.noSC.log ~ age.log +
        age.cat + stype + sex))[,-1] )
head(diss.model)
yName <- "W.noSC.log"
xName <- colnames(diss.model)[-1]
xName
# run model
mcmc.dm2 <- genMCMC(data=diss.model, xName=xName, yName=yName,
        numSavedSteps=numSavedSteps, thinSteps=thinSteps,
        saveName=fileNameRoot, nChains=nChains)
str(mcmc.dm2)
# extract infos
mcmc.red <- as.mcmc(lapply(mcmc.dm2, function(i) i[,c(1:8,17)]))
color_scheme_set("pink")
mcmc_parcoord(mcmc.red, pars=colnames(mcmc.red[[1]]))
mcmc_parcoord(mcmc.red, pars=colnames(mcmc.red[[1]])[-c(9)])
color_scheme_set("brightblue")
# scale before
mcmc_parcoord(mcmc.red, pars=colnames(mcmc.red[[1]]),
        transform = function(x) {(x - mean(x)) / sd(x)})
mcmc_parcoord(mcmc.red, pars=colnames(mcmc.red[[1]])[-c(9)],
        transform = function(x) {(x - mean(x)) / sd(x)})
```

Otros gráficos (no impresos) son, por ejemplo

```
# just sigma and nu
mcmc_parcoord(mcmc.red, pars=colnames(mcmc.red[[1]])[c(8,9)])
mcmc_pairs(mcmc.red)
# selected beta[1] vs. nu
mcmc_scatter(mcmc.red, pars=c("beta[1]","nu"),
        transform=list(nu="log"))
# traceplot MCMCs
mcmc_trace(mcmc.red)
# autocorrelation MCMCs
mcmc_acf(mcmc.red, lags=10)
```

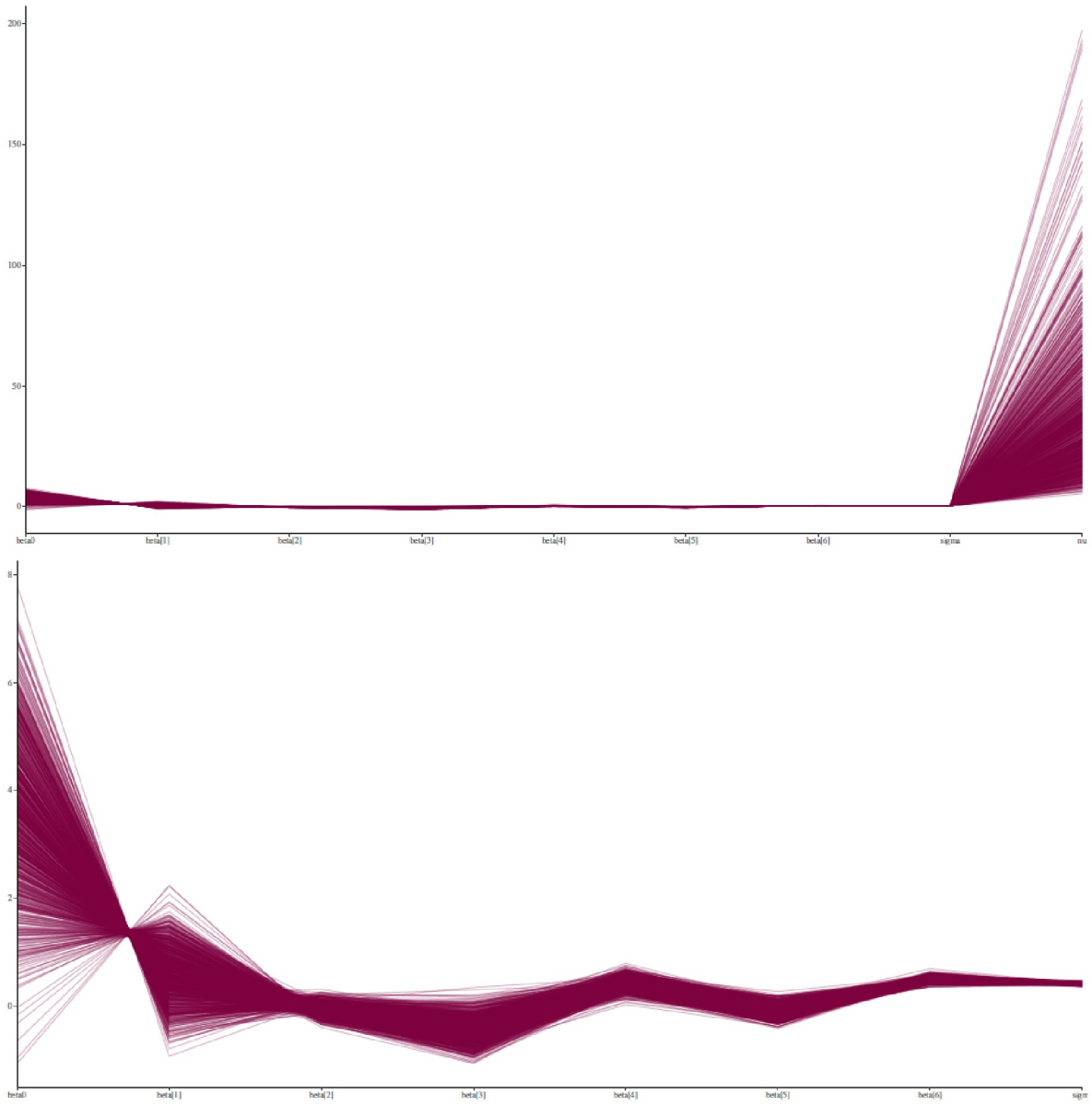


Figura 6.61. *Estudio de Gürtler (2005, humor – producción de palabras, análisis MCMC no escalado/escalado, parte 1).*

Para las comprobaciones predictivas posteriores, se extraen las cadenas MCMC y se pueden buscar los coeficientes mediante expresiones regulares.

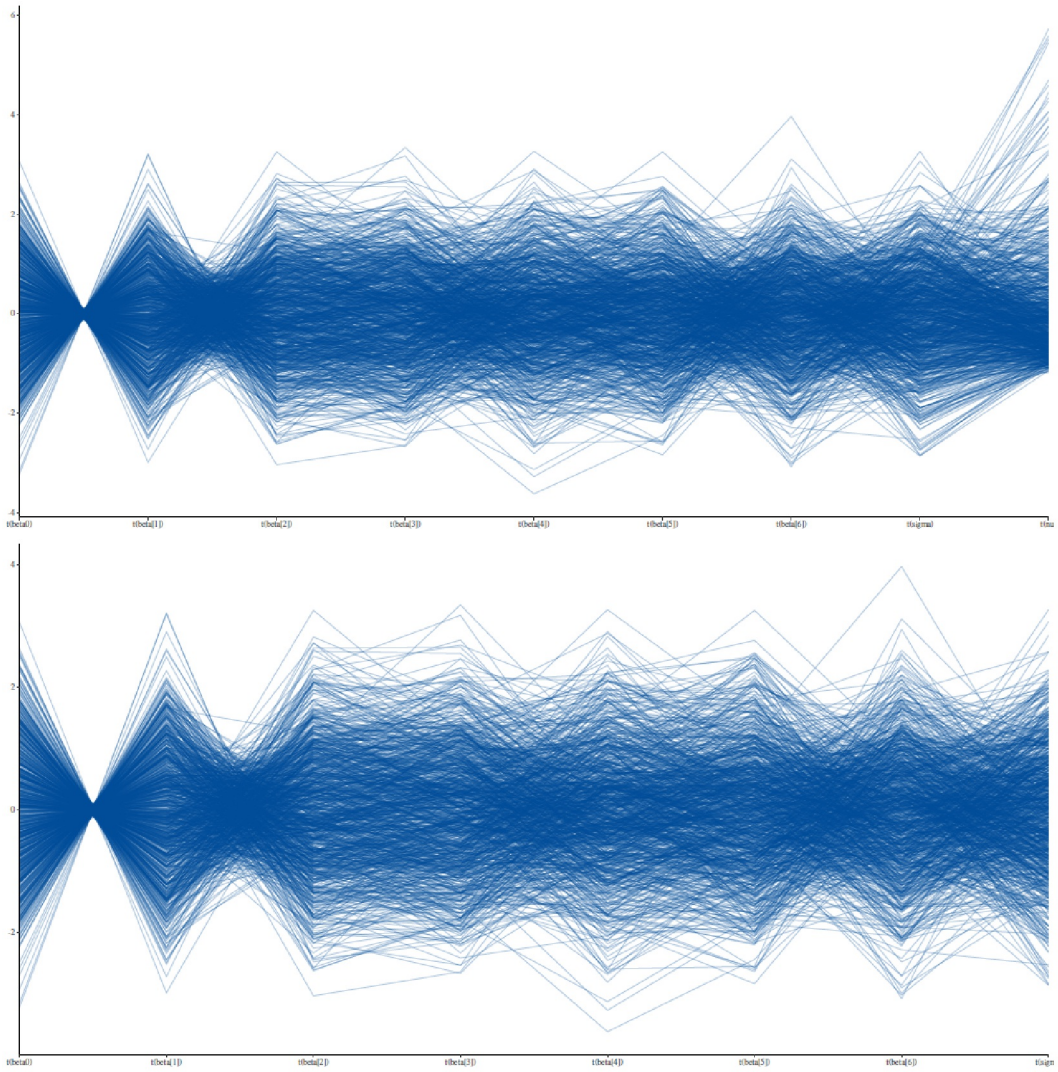


Figura 6.62. *Estudio de Gürtler (2005, humor – producción de palabras, análisis MCMC no escalado/escalado, parte 2).*

```
> mcmc.mat.dm2 <- as.matrix(mcmc.dm2)
> d.mcmc.mat.dm2 <- dim(mcmc.mat.dm2)
> head(mcmc.mat.dm2[,grep("^beta",colnames(mcmc.mat.dm2))])
      beta0  beta[1]  beta[2]    beta[3]  beta[4]  beta[5]  beta[6]
[1,] 2.5928  0.93303 -0.145933  -0.633656 0.46048 -0.082602 0.48882
[2,] 4.6020  0.17558  0.035108  -0.243898 0.41318 -0.092431 0.48851
[3,] 3.4895  0.52838  0.042177  -0.431131 0.54487  0.065936 0.53412
[4,] 4.0754  0.35067  0.033827  -0.208606 0.39827 -0.069973 0.56544
[5,] 2.4294  0.90771  0.031069  -0.415738 0.48452  0.086913 0.46387
[6,] 5.8112 -0.34216  0.132703   0.081163 0.43405  0.028801 0.68995
> beta.nams <- grep("^beta",colnames(mcmc.mat.dm2),value=TRUE)
> beta.id <- grep("^beta",colnames(mcmc.mat.dm2))
> beta.nams
[1] "beta0" "beta[1]" "beta[2]" "beta[3]"
[5] "beta[4]" "beta[5]" "beta[6]"
> beta.id
[1] 1 2 3 4 5 6 7
```

Hay dos tipos de gráficos en `mcmc.diag.kruschke()` que permiten la salida a archivos si es necesario. Para ello, el parámetro `saveName` debe estar establecido y no ser cero. La opción `PLOTmult=TRUE` devuelve gráficos del MCMC, la autocorrelación, la última cadena MCMC con factor de reducción ("shrink factor") y un gráfico de estimación de densidad con HDI, mientras que `PLOThist=TRUE` produce histogramas y un diagrama de pares a través de todos los parámetros.

Ahora hay que decidir sobre qué base se van a realizar las comprobaciones predictivas posteriores. Por un lado, es posible simular para casos individuales o combinaciones de valores (edad, grupo de edad, tipo de escuela, sexo) o seleccionar aleatoriamente casos enteros a partir de los valores empíricos existentes para generar una muestra correspondiente a la empírica y simular sobre ella. Del mismo modo, se pueden calcular las medias y las desviaciones estándar a lo largo de las Posteriores para generar valores aleatorios a partir de la distribución normal con estos valores, a saber tantas veces como se deseen simulaciones predictivas (Clark, 2018, capítulo "Model Checking"). Elegimos la opción de muestreo de casos completos y más adelante mostramos cómo se pueden implementar análisis dirigidos de valores concretos o cómo extraer de la distribución normal. Si solo se lanzaran al azar parámetros de los datos empíricos, es decir, si se generaran casos mediante bootstrap a partir de los datos empíricos sin estructura inherente, el modelo existente dejaría de ser lo suficientemente apropiado para el ejemplo. Reproducir esto sería tarea de lectores dedicados. Preliminarmente, el bootstrap sobre la base de casos enteros o extrayendo de la distribución basada en los valores de las Posteriores no contradice el modelo. Lo que todas estas opciones tienen en común es que generan una cierta cantidad de ruido – es decir, una variación de los datos empíricos – y, por tanto, una cierta cantidad de incertidumbre en el sentido del bootstrap, que es típica de los datos nuevos y necesaria para probar la coherencia y aplicabilidad del modelo existente con respecto a las nuevas variantes de datos.

En primer lugar, sin embargo, pasamos al procedimiento técnico y creamos una muestra bootstrap a partir de la muestra empírica existente con `sample()` (véase el capítulo 4.3.5) para muestrear casos completos. A continuación, definimos una matriz cuyas filas son tan grandes como la cadena MCMC y las columnas corresponden al tamaño de la muestra.

```
# create a random sample (full cases!) from real values
# sample with replacement = bootstrap
seed <- 1432
set.seed(seed)
head(diss.model)
diss.model.d <- dim(diss.model)
diss.model.d #360 7 > > > N=360 persons (rows), k=7 variables (cols)
samp.ids <- sample(1:diss.model.d[1], replace=TRUE)
samp.mat <- diss.model[samp.ids,]
d.samp.mat <- dim(samp.mat)
d.samp.mat
# create y_pred values from posterior values
# for each of the sample elements
mat.ypred <- matrix(data=NA, ncol=d.samp.mat[1],
       nrow=d.mcmc.mat.dm2[1])
# cols=360 Persons
# rows=1002 mcmc (=3*334)
# ie. for each person investigated
# from the sample a full mcmc chain
```

Una pequeña multiplicación matricial genera ahora las predicciones Y^{pred} basadas en el modelo, ya que la multiplicación de matrices combina las cadenas MCMC (estimaciones de parámetros, es decir, β-coeficientes de las Posteriores) con los valores de salida muestreados (como un nuevo conjunto de datos).

```
# we use only the betas R-Code
attr(mcmc.mat.dm2,"dimnames")
dim(mat.ypred)
for(i in 1:d.samp.mat[1])
{
```

```
mat.ypred[,i] <- mcmc.mat.dm2[,beta.id] %*% c(1,t(samp.mat[i,-c(1)]))
}
```

Esto resulta en un valor esperado μ^{pred} para la producción de palabras por sorteo MCMC. Así, para cada extracción MCMC, la ecuación del modelo se calcula completamente sobre la base de los valores iniciales muestreados. Esto puede hacerse de forma resumida para toda la muestra o para cada caso individual. Depende del interés de la investigación. Calculamos el valor medio por individuo simulado.

```
> muPred <- apply(mat.ypred,1,mean)
> mean(muPred)
[1] 5.268682
> sd(muPred)
[1] 0.0224466
> head(muPred)
[1] 5.304599 5.258361 5.274514 5.264946 5.249873 5.237704
> length(muPred)
[1] 1002
```

A continuación se realiza la predicción basada en el modelo, pero según el código R de Kruschke (2016d) enriquecido con algo de ruido adicional de una distribución *t* con el número de grados de libertad, que corresponde al parámetro ν de los datos de las Posteriores. Esto se debe a que la distribución predictiva posterior de la distribución normal es una distribución *t* (Murphy, 2007; Wikipedia, 2019e para una tabulación de otras distribuciones) cuando no se conoce la varianza. La distribución predictiva posterior tiene la misma media que la Posterior, pero una varianza mayor porque añade ruido a la distribución debido a la simulación (predicción).

Esto es similar al procedimiento de imputación múltiple (Yan, 2016-02), ya que este procedimiento conduce idealmente a una media constante y una varianza mayor. Este ruido se añade al ruido anterior mediante el muestreo de los casos, añadiendo más ruido basado en la distribución *t*, que es más robusta en las colas (= extremos de la distribución) que la distribución normal. En estadística clásica, esto significa que los valores *p* no son tan extremos.

Así pues, el valor predictivo Y^{pred} surge de la media simulada μ^{pred} más el producto de la varianza residual `sigma` por valores aleatorios de la distribución *t* con parámetro ν.

```
# Y_rep is a function of
# mu_pred + sigma +
# noise-due-to-prediction-here-based-on-t-dist-and-nu
# (=df-of-t)
Y_rep <- muPred + mcmc.mat.dm2[,"sigma"] *
rt(nrow(mcmc.mat.dm2),df=mcmc.mat.dm2[,"nu"])
```

Se puede trazar la distribución predictiva posterior (véase la Fig. 6.63) y marcar en ella los valores empíricos (es decir, la media) para detectar gráficamente las desviaciones de los supuestos del modelo.

```
# plot
hist(Y_rep, prob=TRUE, pre.plot=grid())
lines(density(diss.model[,"W.noSC.log"]), col="darkred", lwd=2)
lines(density(Y_rep), lwd=2, col="blue")
```

En primer lugar, algunas estadísticas resumidas de Y_rep y muPred:

```
> # summary statistics
> lapply(list(muPred=muPred,Y_rep=Y_rep),
function(x) c(summary(x),sd=sd(x)))
$muPred
Min. 1st Qu. Median Mean 3rd Qu. Max. sd
```

```
5.1973818 5.2520000 5.2691751 5.2686820 5.2839847 5.3389869 0.0224466
$Y_rep
Min. 1st Qu. Median Mean 3rd Qu. Max. sd
3.9541239 5.0130603 5.2881143 5.2779998 5.5460456 6.5746025 0.4263025
> # difference
> mean(Y_rep-muPred)
[1] 0.00931784
> # effect size
> mean(Y_rep-muPred)/sd(Y_rep)
[1] 0.02185734
```

El valor *p* bayesiano como variante numérica de la comprobación del modelo resulta de la masa de la distribución de Y_rep > muPred. A continuación, comparamos Y_rep con los datos empíricos.

```
> # Bayesian p-value
> # compare predict vs. predictive
> 1-mean(Y_rep > muPred)
[1] 0.4830339
> # compare predictive vs. empirical
> 1-mean(Y_rep > mean(diss.model[,"W.noSC.log"]))
[1] 0.4870259
```

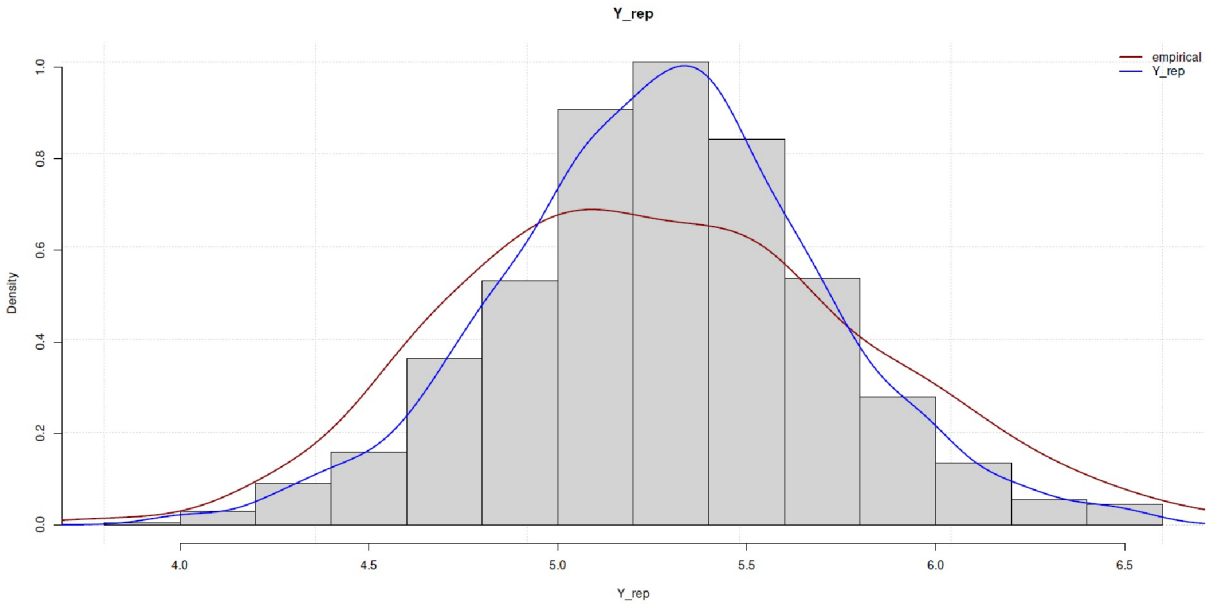


Figura 6.63. *Estudio de Gürtler (2005, humor – producción de palabras, distribución predictiva posterior).*

En este caso, no hay razón aparente para suponer una inadaptación bruta del modelo. El modelo parece captar y describir suficientemente bien la estructura de los datos; nadie piensa que esto muestre la verdadera distribución o el verdadero modelo. Esto haría que el modelo cuasi *t* pudiera utilizarse para nuevos datos – bajo las condiciones de las extensiones anteriores como rellenar las lagunas debidas al muestreo truncado. No obstante, se observa que los datos empíricos parecen ser más amplios, lo que deja margen para mejorar el modelo, ya que no todas las áreas de datos están cubiertas por igual.

Volvamos a las variaciones mencionadas sobre cómo se pueden generar otros datos de origen para crear la distribución predictiva posterior. En lugar de muestrear casos enteros, se podrían crear nuevos casos a

partir de valores empíricos aleatorios combinando los valores individuales de diferentes unidades de estudio (= personas). La llamada sería entonces la siguiente (extracto del script de R):

```
mat.ypred <- matrix(data=NA, ncol=dim(diss.model)[1],
       nrow=d.mcmc.mat.dm2[1])
Y_rep.mat <- matrix(data=NA, nrow=dim(mat.ypred)[1],
       ncol=anzsim.reps)
mat.ypred.d <- dim(mat.ypred)
Y_rep.mat.d <- dim(Y_rep.mat)
diss.model.d <- dim(diss.model)
# without response # cols of original table with data
# c(1:6)
varsofinterest <- c(2,3,4,5,6,7)
# -1 because one is the response
anzbetas <- diss.model.d[2]-1
# create random sample from random draws
# from real values, but not full cases
# not full cases, but categories ie. values
for(j in 1:mat.ypred.d[2])
{
case.ids <- sample(1:diss.model.d[1], anzbetas, replace=TRUE)
samp.mat[j,] <- diag(diss.model[case.ids,varsofinterest])
}
```

El resto es similar al procedimiento anterior. Lógicamente, esto lleva a una incertidumbre algo mayor en la predicción que sacar casos enteros, ya que aquí se mezclan valores para los que no está claro si (pueden) ocurrir empíricamente de esta manera.

Otro procedimiento toma la matriz MCMC estimada (= posterior) y simula valores distribuidos de manera normal para cada fila sobre la base de sus valores medios y desviaciones estándar al nivel del tamaño de la muestra. Para ello, primero se resuelve la ecuación del modelo para cada fila con los valores de los parámetros estimados, es decir, se predicen los valores a partir de las estimaciones.

```
varsofinterest # not response (= 1)
str(diss.model)
model.pred <- mcmc.mat.dm2[,beta.id] %*%
t( data.frame(1,diss.model[,varsofinterest]))
str(model.pred)
dim(model.pred)
```

A partir de ahí, se puede obtener el valor medio y la desviación típica por fila.

```
MW <- apply(model.pred,1,mean)
SD <- apply(model.pred,1,sd)
```

Los valores así obtenidos se utilizan para simular valores de la distribución normal. El número de simulaciones puede determinarse libremente. Para cada simulación, se calcula el valor medio y se enriquece con el producto de la varianza residual `sigma` y valores aleatorios de la distribución *t* con el parámetro `v`. El resto del procedimiento es el mismo que el anterior: trazar, valor *p* bayesiano, etc.

```
anzsim.reps <- 100
Y_rep.mat.1 <- sapply(1:anzsim.reps,
       function(s) rnorm(n=d.mcmc.mat.dm2[1], mean=MW[s], sd=SD[s]))
dim(Y_rep.mat.1)
Y_rep.1 <- apply(Y_rep.mat.1,1,mean) + mcmc.mat.dm2[,"sigma"] *
       rt(nrow(mcmc.mat.dm2),df=mcmc.mat.dm2[,"nu"])
```

Para concluir este procedimiento, mostramos cómo se pueden simular valores para una sola persona. Se establece una combinación de parámetros como así: "edad 15 año"s `age.log=log(15)`, parte de la categoría "edad entre [LY 17-LY 19]" `age.cat(17-19]=1`, no forma parte de la categoría "edad entre [LY 20-LY 25]" `age.cat(20-25]=0` y "Gymnasium" `stypeG=1`, no "Realschule" `stypeR=0` y sexo femenino `w=1`.

```
# muPred for single cases
# selected case
dm.nams <- colnames(diss.model)[varsofinterest]
# 15 years, associated age category, stypeG=1, stypeR=0, sexw=1
mat.case1 <- matrix(data=c(1,log(15),1,0,1,0,1), ncol=1,
       dimnames=list(c("Intercept",dm.nams)))
```

A continuación se calcula la ecuación del modelo utilizando la posterior.

```
muPred.s <- mcmc.mat.dm2[,beta.id] %*% mat.case1
```

y se puede seguir procesando como de costumbre. Alternativamente, la simulación de un solo caso con

```
# different case
mat.case1 <- matrix(data=c(1,log(17),1,0,0,1,0), ncol=1,
       dimnames=list(c("Intercept",dm.nams)))
# random sample n=1
muPred.s.rn <- mcmc.mat.dm2[,beta.id] %*%
c(1,t(diss.model[sample(1:d.samp.mat[1],1),-c(1)]))
# calculate prediction based on predictors + sigma + rt noise
set.seed(1823)
Y_rep.s <- muPred.s.rn + mcmc.mat.dm2[,"sigma"] *
       rt(nrow(mcmc.mat.dm2),df=mcmc.mat.dm2[,"nu"])
```

Ahora se podría investigar cómo grande resulta la incertidumbre introducida debida al ruido de la distribución *t*.

```
> mean(muPred.s)
[1] 5.957871
> mean(Y_rep.s)
[1] 5.483334
> mean(muPred.s.rn)
[1] 5.488022
> mean(Y_rep.s-muPred.s.rn)
[1] -0.004687908
> mean(Y_rep.s-muPred.s.rn)/sd(Y_rep.s)
[1] -0.0106433
```

Sería ir demasiado lejos profundizar en la modelización y las comprobaciones predictivas posteriores asociadas. Se puede encontrar fácilmente más bibliografía sobre el tema, ya sea a nivel conceptual (Gelman, Carlin, Stern & Rubin, 2004; Gelman, Hwang & Vehtari, 2013; Kruschke, 2013c) o en forma de tutoriales (Muth, Oravecz & Gabry, 2018; Gabry & Goodrich, 2018), así como entradas de blog, debates y código de R (Gelman, 2007b, 2009a; Kruschke, 2016c, 2017b). Los paquetes de R `bayesplot`, `rstantools`, `rstanarm` y `brms` tienen herramientas gráficas como `pp_check()` y funciones como `posterior_predict()`, que permiten examinar más de cerca los modelos existentes con datos simulados para determinar su valor predictivo o aplicarlos a nuevos datos empíricos, porque técnicamente es lo mismo. En la interpretación, por supuesto, los casos difieren.

6.9. Replicación y teorema de Bayes

En el contexto de los argumentos a favor de la necesidad de las réplicas ya desarrollados en el capítulo 4.4.4, el teorema de Bayes permite explícitamente que el conocimiento conocido y, por tanto, las ideas sobre parámetros u otros resultados de estudios se incorporen como información previa (la Prior) para futuros diseños de investigación (véase el estudio de caso de las tasas de aprobados en la terapia de adicción con hospitalización, capítulo 6.15.2). Así, se puede entender el teorema de Bayes como un ciclo de auto-actualización, de modo que, en términos de saturación de información, la replicación puede continuar teóricamente hasta que las estimaciones de los parámetros se estabilicen, o hasta que surja un nuevo modelo. Marsman, Schönbrodt, Morey, Yao, Gelman y Wagenmakers (2017) utilizan datos disponibles públicamente para mostrar cómo se pueden implementar y discutir una serie de estudios clásicos de replicación psicológica social (Nosek & Lakens, 2014) utilizando la estadística bayesiana.

Comparable al procedimiento Neyman-Pearson (análisis de potencia a priori, véanse los capítulos 4.3.3.1 y 4.4.3, respectivamente), se puede utilizar la estadística de Bayes también para diseñar y planificar estudios (Schönbrodt & Wagenmakers, 2018). Los autores recomiendan una Prior más amplia para la planificación y una Prior más escéptica y estrecha para el análisis concreto de los datos. Estos análisis se realizan con la ayuda de tamaños de efecto anticipados y factores de Bayes ("Bayes Factor Design Analysis", BFDA, ibíd.; Stefan, Gronau, Schönbrodt &Wagenmakers, 2019; paquete R BFDA). Al hacerlo, los autores elaboran tres clases diferentes de diseños:

1. Diseños *N* fijos con un tamaño de muestra fijo *N*

2. Diseños abiertos con factor Bayes secuencial (SBF), en los que las pruebas se realizan después de cada participante en el estudio y los datos se recopilan hasta que hay pruebas claras a favor de las hipótesis H_0 o H_1. Cabe recordar aquí que el estadístico de Bayes no aborda el problema de las pruebas múltiples del mismo tipo o "p-hacking" (véase el capítulo 4.4.2). Wagenmakers (2007a, 2007b) describe un enfoque adecuado de las reglas de parada, pero se centra en las pruebas más que en la estimación.

3.Una versión modificada del SBF, en la que los datos se recogen hasta un tamaño de muestra máximo, independientemente de si en este punto ya se dispone de pruebas claras y convincentes para H_0 o H_1. La elección del enfoque debe venir determinada por la pregunta de investigación.

6.10. Deducción frente a inducción

Jaynes (2003, p.311), en su libro sobre la teoría de la probabilidad como lógica, define el propio significado de inducción como

> „On the other hand; if the predictions prove to be wrong, then induction has served its real purpose; we have learned that our hypotheses are wrong or incomplete, and from the nature of the error we have a clue as to how they might be improved."

Esta opinión probablemente la compartiría el estadístico Andrew Gelman, que recomienda la prueba gráfica de modelos para comprender dónde un modelo no puede explicar los datos con el fin de aprender de él para obtener un modelo mejor.

Aunque las matemáticas de la estadística bayesiana no son o no pueden ser cuestionadas, sí lo es su interpretación (Hammerton, 1968; Jaynes, 2003). Sin embargo, se trata de un problema epistemológico, tanto

que incluso filósofos como Popper lo comentaron (1943, Cap. XVII), por lo que repetimos algunos argumentos del Capítulo 1.2 sobre epistemología y filosofía de la ciencia y los relacionamos con la estadística bayesiana. Por desgracia, Popper pasó por alto en su argumentación que la estadística de Bayes no busca una probabilidad absoluta de una posible hipótesis o teoría. Más bien se compara la probabilidad posterior de una hipótesis con otras hipótesis – y así se obtiene a partir del teorema de Bayes un poder explicativo relativamente mayor de una hipótesis en un *contexto finito y no* en comparación con todas las hipótesis potencialmente posibles cuyo número se aproxima al infinito. En realidad, las hipótesis en liza se limitan a unas pocas contables. En consecuencia, los cálculos de Popper (ibíd.) son erróneos porque no se ajustan a la situación. El enfoque bayesiano es bastante coherente con un enfoque falsacionista (Gelman, 2011a) y es entonces cuando tiene lugar la comprobación de modelos (Gelman, Carlin, Stern & Rubin, 2004, cap. 6). También es posible trabajar inductivamente con la estadística de Bayes, probar modelos deductivamente y luego comprobarlos o modificarlos, que es lo que deberían mostrar las explicaciones anteriores sobre, entre otras cosas, los posterior predictive checks (véase el capítulo 6.8.4.3). Popper consideraba que la inducción era imposible para la generación de conocimiento y negaba que fuera una forma de conocimiento por derecho propio. Interesantes de leer son los contraargumentos de Cox (1961, p.91.), que en realidad están dirigidos a Hume, pero que encajan bien con Popper. Hume argumentaba de forma diferente, describiendo la inducción más bien como un difícil subproducto de la naturaleza humana y la consideraba no racional, es decir, sin fundamento lógico (Greenland, 1998 para más detalles), sino como existente. Neyman, por su parte, asumía que todas las inferencias son de naturaleza deductiva. Todas estas posiciones son criticadas ferozmentes por Jaynes (2003, p.276., p.310., p.499), representante de la dirección objetiva de Bayes (ibid., p.310),

> „Philosophers have argued over the nature of induction for centuries. Some, from David Hume (1711–76) in the mid-18th century to Karl Popper in the mid-20th (for example, Popper and Miller, 1983), have tried to deny the possibility of induction, although all scientific knowledge has been obtained by induction."

Y otra vez (ibíd., p.499),

> „Fisher and Jeffreys, aware that all scientific knowledge has been obtained by inductive reasoning from observed facts, naturally enough denied the claim of Neyman that inference does not use induction, and of the philosopher Karl Popper that induction was impossible."

Neyman (1950) estaba preparado para hablar de *comportamiento inductivo* (véase el capítulo 4.3.3), a saber de comportamiento dirigido por datos basado en el resultado de una prueba. Fisher (1955), en cambio, habló directamente de *razonamiento inductivo* y criticó la actitud de Neyman y la de los lógicos (ibíd., p.74),

> „Logicians, in introducing the terms 'inductive reasoning' and 'inductive inference' evidently imply that they are speaking of processes of the mind falling to some extent outside those of which a full account can be given in terms of the traditional deductive reasoning of formal logic. Deductive reasoning in particular supplies no essentially new knowledge, but merely reveals or unfolds the implications of the axiomatic basis adopted. [...] It is the function of inductive reasoning to be used, in conjunction with observational data, to add new elements to our theoretical knowledge."

La *inducción matemática* se denota como:

Recordatorio 6.3: Inducción matemática

Si se puede demostrar que algo que es cierto para n también lo es para $n + 1$, esto se llama inducción matemática.

La inducción matemática sólo puede aplicarse en abstracto. En la realidad empírica no existe tal inducción absoluta; y esto es una realización trivial. No obstante, tales inferencias inductivas generalizadas deben aplicarse en la práctica de la investigación de forma no restringida y no necesariamente en la estadística bayesiana (Meehl, 1990). Las cosas cambian cambian continuamente; y el conocimiento, como parte del mundo, siempre cambia con él y, por lo tanto, debe ser (re)descubierto para ser examinado críticamente a la luz de la realidad. Tukey (1977, 1980) vio esto de forma similar, comentando que se prueban estadísticamente demasiados modelos de modo confirmatorio en lugar de desarrollar hipótesis y teorías fructíferas sobre los datos. Tukey cree que a veces encontrar las preguntas adecuadas es más importante que encontrar las respuestas, por lo que el análisis exploratorio de datos refleja una actitud y no simplemente un conjunto de técnicas. Eso es más difícil que el análisis confirmatorio, pero que, en cabio, es más fácil de enseñar. Sin embargo, ambos son necesarios para progresar: Inducción y Deducción.

Como señala Gelman (2011a), la estadística de Bayes no es en absoluto contraria a la necesidad de falsar. Más bien al contrario. Las comprobaciones predictivas posteriores y pruebas gráficas de modelos descritas anteriormente implementan exactamente lo que se necesita para probar, mejorar y adaptar modelos, para mejorarlos y adaptarlos a situaciones de información nuevas y cambiadas. La propia estructura del teorema de Bayes significa que la nueva información puede combinarse con los conocimientos existentes para llegar a un estado de conocimiento modificado. El teorema de Bayes, gracias a su posibilidad de actualización, permite aprender de la experiencia (véase el capítulo 6.3.2.5 para un estudio de caso discreto). Los factores de Bayes persiguen el mismo objetivo con diferentes medios. Un artículo que merece la pena leer sobre cómo se deberían manejar realmente las hipótesis e inferencias probabilísticas es el de Greenland (1998). El autor ve el problema en el dominio semántico, el correspondiente uso de conceptos y, en consecuencia, la errónea recepción de Popper y otros autores. Conceptos clave como inducción se utilizan muy diferentemente en distintas épocas y por distintos autores , por lo que puede surgir confusión en cuanto a qué se entiende por qué y sobre qué – en realidad – estamos discutiendo. El acuerdo exacto sobre los conceptos es un requisito previo para todas las discusiones posteriores. Greenland da ejemplos de ello (ibid., p.545).

En cualquier caso, la cuestión de la objetividad y la subjetividad no es tan fácil de decidir, como demuestran, por ejemplo, las observaciones de McElreath (2015, p.10) sobre la falsificación. Falsificación en sí misma se basa en normas consensuadas en una comunidad científica, lo que generalmente se considera como evidencia aceptable. La falsificación a lo largo de los datos empíricos no es de naturaleza lógica sino consensual, a saber, cuál es el significado de la evidencia empírica. Dentro de la comunidad científica esto apenas se nota, porque todo el mundo argumenta de esta manera. Hay que dar un paso atrás y reflexionar sobre cuáles son realmente los principios del propio trabajo. Estos significados consensuados no son siempre fijos, sino que cambian con el tiempo. Como ejemplos conocidos, cabe citar el peligro de la radiactividad o los rayos X, operacionalizados por sus valores límite. A finales del siglo XIX todavía se realizaban experimentos sin protección. Lo mismo ocurría con los mineros que extraían material radiactivo – sin protección. En consecuencia, al principio los radiólogos morían y los mineros cayeron gravemente enfermos, etc. Lo mismo ocurre con los relojes cuyas joyas contenían algo de radio, que brillaban por la noche. También ellos eran nocivos y a la larga fatales para los pilotos que gustaban de utilizarlos. En los años 30 era incluso chic tomar compuestos de radio para diversas enfermedades, sin ningún tipo de precaución. Un punto de inflexión llegó con la devastadora de las bombas atómicas lanzadas sobre Hiroshima y Nagasaki al final de la Segunda Guerra Mundial. A ello se sumaron las pruebas con bombas atómicas y bombas de hidrógeno, que causaron enormes daños, por ejemplo en el Atolón de Bikini -- y luego la Guerra Fría entre Oriente y Occidente. Así, poco a poco los mecanismos de la radiación y su efecto en los sistemas biológicos se fueron aclarando, y en consecuencia se pudieron ajustar los límites de lo aceptable, es decir, endurecerse drásticamente. Una falsificación en diferentes momentos de la historia sobre si la radiactividad es nociva o no habría conducido a resultados diferentes y a las correspondientes recomendaciones. Esto ocurre en cada caso en función de los conocimientos y las posibilidades técnicas del momento. Y no sabemos cómo se perfeccionarán y modificarán los resultados en el futuro. El final está aún muy lejos, y otras cuestiones como la radiación de los smartphones, las antenas de telefonía móvil, las redes WLAN, etc. están lejos de ser investigadas a fondo en lo que respecta a sus efectos sobre los seres humanos y los sistemas biológicos. Por tanto, la falsificación depende del momento y del contexto. No hay lugar para afirmaciones generales atemporales.

Recordatorio 6.4: Muestras finitas

Afortunadamente, las afirmaciones de la estadística de Bayes siguen siendo adecuadas al contexto y están relacionadas con muestras finitas, que consideramos muy adecuadas a la realidad. En cambio, la estadística clásica intenta aplicar condiciones asintóticas en sus supuestos que son difíciles o imposibles de justificar en la práctica.

Así, el teorema de Bayes permite extraer conclusiones ante una información contextual finita. Una inferencia inductiva de cualquier tipo a condiciones infinitas no tiene lugar y no se pretende. La estadística de Bayes es interesante precisamente, porque permite ante información incierta y limitada (Studer, 1996b), sacar conclusiones plausibles utilizando toda la información disponible – una característica que la estadística frecuentista no puede ofrecer. Allí se tiene utilizar la inferencia de la teoría Neyman-Pearson a condiciones de infinito (asintótico). Sin embargo, sería erróneo considerar la estadística clásica como puramente inductiva, ya que el marco (véase la teoría de Neyman-Pearson, capítulo 4.3.3) es hipotético-deductivo, aunque las decisiones basadas en datos puedan tomarse de forma inductiva. Esto es diferente en el caso de Fisher, ya que él se ocupa del conocimiento inductivo.

Para la estadística de Bayes, la pretensión de una inferencia inductiva siempre válida no se hace ni se pretende, como subraya Jaynes (2003, p.310, cursiva en el original):

> „In denying the possibility of induction, Popper holds that theories can never attain a high probability. But this presupposes that the theory is being tested against an infinite number of alternatives. We would observe that the number of atoms in the known universe is finite; so also, therefore, is the amount of paper and ink available to write alternative theries. It is not the absolute status of an hypothesis embedded in the universe of all conceivable theories, but the plausibility of an hypothesis relative to a definite set of specified alternatives, that Bayesian inference determines."

Es mucho más cierto, sin embargo, que las afirmaciones empíricamente fundadas relacionadas con la realidad limitada no son o no pueden ser nunca afirmaciones totales, ni en su formulación positiva ni en su formulación negativa. Para los siempre populares cisnes en sus diversos colores, esto no significa, como McElreath (2015, p.9) señala:

> *Todos los cisnes son blancos.*

En cambio, la hipótesis es

> H_0: *El 80 % de los cisnes son blancos.*

o también

> H_0: *Los cisnes negros son raros.*

Inmediatamente se hace evidente que una hipótesis de este tipo no puede responderse simplemente con un *sí* o un *no*, sino que se deben tener en cuenta las zonas grises y la incertidumbre como factores efectivos (véase el capítulo 6.8.4.2 sobre el concepto de ROPE como inspiración). La hipótesis nula en sí no afirma que no haya cisnes negros, sino sólo que tienen una cierta frecuencia. La tarea (o a veces el problema) consiste *tanto en estimar como en explicar* la distribución de los colores de los cisnes. La simple aplicación del modo

Tollens (véase la tabla 2.6, p.25) con la ayuda de simples afirmaciones "si A entonces B" carece, por tanto, de sentido. Ahora cabría preguntarse si tal hipótesis es científica. McElreath (2015) ve esto de tal manera que no es tanto la cuestión de la cientificidad lo que interesa, sino más bien el hecho de que las cuestiones que interesan en la realidad a menudo corresponden precisamente a estas formulaciones y como tales *se les da relevancia*. Son las afirmaciones basadas en la probabilidad y justificadas contextualmente que son significativas para las preguntas concretas y no las simples respuestas globales, aunque estas últimas serían, por supuesto, más agradables.

Por ejemplo, podríamos preguntarnos si el asesoramiento organizativo en las escuelas tiene un efecto demostrable. Ya podemos responder con un alto grado de certeza: "¡Por supuesto!", porque cualquier otra cosa tendría que estar justificada, es decir, que una intervención externa no tiene ningún efecto. Esto equivaldría a demostrar un efecto nulo (véase el capítulo 4.4.9 sobre procedimientos equivalentes), lo que en la práctica (véase la paradoja de Meehl, capítulo 4.4.14.3) no es tan fácil o incluso imposible. Pero incluso en ese caso, la ganancia de conocimiento sigue siendo muy modesta. Sólo cuando consideramos las condiciones concretas (por ejemplo, el tipo de escuela, la clarificación de la tarea, el problema, la historia previa, las personas implicadas, la duración, etc.) en relación con diferentes áreas problemáticas (por ejemplo, problemas internos en el colegio, coordinación entre la dirección y el colegio, trabajo con los padres, etc.) y obstáculos (p. ej. falta de voluntad de cambio, dinero + personal, presión de tiempo, múltiples exigencias, etc.), se pueden plantear preguntas que merece la pena explorar. Sin embargo, las afirmaciones resultantes son entonces de naturaleza compleja y dejan de ser triviales. Tendríamos que complementarlas con un cálculo de pérdidas de lo elevado que es el esfuerzo (normalmente operacionalizado a través del dinero) en comparación con el beneficio, y habría que especificar el propio beneficio, en el que no sólo influyen factores monetarios, sino también políticos, sociales y de otro tipo. Por eso se necesitan modelos complejos y, probablemente, enfoques de investigación multimétodo que integren todos los tipos de datos pertinentes (métodos mixtos). La estadística bayesiana puede desempeñar un valioso papel en este proceso.

Al igual que con la cuestión de lo objetivo frente a lo subjetivo, en este punto la discusión parece girar en torno a algo que de todos modos no se puede resolver. El razonamiento lógico en la abstracción de la lógica y sus silogismos puede definirse con relativa claridad. Enfrentado a datos concretos en un contexto real, ya no parece tan sencillo. La complejidad es mucho mayor. Las interpretaciones son – aunque haya acuerdo sobre el color de los cisnes blancos y negros (véase más arriba, McElreath, 2015, capítulo 1.2) – no independientes entre sí. Además, siempre hay que preguntarse a qué fase del proceso de investigación empírica se refieren las cuestiones de inducción y deducción – la generación de teorías, el establecimiento de modelos, la comprobación de modelos mediante análisis de datos, la traslación a la práctica, etc.

En el contexto de AED (véase el capítulo 5) para identificar estructuras y patrones, se fomenta mucho el elemento inductivo. En cambio, en el contexto de un experimento de prueba, el elemento deductivo pasa a primer plano. Uno no hace que el otro sea más o menos científico, ya que las preguntas de investigación son completamente diferentes. Si la investigación se observa durante un periodo de tiempo más largo, se puede encontrar un ciclo constante de formulación de hipótesis, examen crítico, revisión, etc. (véase la Fig. 2.1, p.8). Sin embargo, no está claro si se produce algún progreso real o si éste es siquiera detectable. En este sentido, podemos inspirarnos en Kuhn (1973) y Feyerabend (1976), que han descrito el desarrollo social de la ciencia (paradigmas) per se y los límites o posibilidades del conocimiento..

A modo de recordatorio, nadie ha conseguido aún trascender la relatividad de los criterios de verdad en la ciencia y desarrollar un criterio absoluto que sea siempre válido y no se derive del consenso discursivo. Mientras esto siga siendo así – e incluso para este supuesto no disponemos de un criterio absoluto – asumimos el cambio del tiempo, es decir, que *el conocimiento es cambiante*. Las cosas vienen, las cosas vuelven. A este respecto, la conclusión lógica para nosotros sería que una conclusión estadística debería referirse a las condiciones reales concretas y no a una inferencia siempre válida en condiciones imaginarias infinitas. En la realidad no existen muestras infinitas. Sin embargo, son mucho más comunes muestras muy pequeñas y una gran cantidad de conocimientos contextuales implícitos que deben conectarse de forma significativa. Desde nuestro punto de vista, la estadística de Bayes puede hacer una contribución inmensamente valiosa, ya que funciona de forma excelente con muestras pequeñas (Studer, 1998; Bretthorst, 1993) y no sólo permite, sino que exige, la inclusión del conocimiento contextual. Esto descartarla, como se tiende a hacer en el contexto de la estadística objetiva de Bayes, nos parece tan engañoso como pretender

prescindir por completo de otras opciones *objetivas* existentes (por ejemplo, estudios previos, pero también consideraciones matemáticas, si son adecuadas al contexto). La *verdad* suele estar en algún punto intermedio.

Desde nuestro punto de vista, toda esta discusión parece un poco la cuestión del huevo y la gallina, a pesar de los argumentos de gente como E.T. Jaynes, Cox, Fisher, Kuhn, Feyerabend, etc., con la mayoría de los cuales podemos estar de acuerdo. En cuanto se acepta, o mejor se cree, que empíricamente no existe un criterio absoluto de verdad y, por tanto, tampoco una evidencia teórica o empírica absoluta, no hay ni inducción ni deducción ni falsación en sentido absoluto. *Siempre podemos equivocarnos*. Quienes creen seriamente que pueden extraer conclusiones generales y universalmente válidas en todos los contextos mediante la inducción – por ejemplo, a través de observaciones – sin contrastarlas críticamente y repetidamente con la realidad, sufren en nuestra opinión una pérdida de realidad al igual que quienes buscan la salvación en diseños exclusivamente deductivos y pasan completamente por alto de dónde surgieron realmente en el pasado y siguen surgiendo las ideas iniciales para tales diseños: de observaciones, experiencias, sucesos inesperados y mucho más. Éstas son de naturaleza inductiva, pero se han desprendido del contexto original de origen a través del debate, la reflexión y otras actividades cognitivas. Las ideas nuevas no salen a la luz (sólo) por deducción el examen crítico no funciona muy bien de forma inductiva. Necesita al menos *ambas cosas*, si no más. Por eso, en el capítulo 2 hemos presentado la *investigación como un ciclo* que intenta reconstruir precisamente estas interdependencias a modo de modelo. Al principio de cualquier discusión de este tipo, los autores deberían, antes de empezar a escribir, consultar a fondo a Bateson (1985) con sus conceptos de *información y puntuación*. La información nos muestra cómo extraer significado subjetivo de las diferencias y la puntuación cómo dividir el flujo continuo del tiempo en bocados digeribles que luego podemos contrastar, hacer predicciones, etc. La puntuación nos muestra cómo extraer significado de las diferencias. Los signos de puntuación señalan que estamos sometidos a una (auto)ilusión constante, a saber de poder separar en nuestra conciencia procesos como la inducción, la abducción y la deducción. Más bien pueden tener lugar en paralelo, pero cada uno de los elementos pasa con más fuerza al primer plano subjetivo de la conciencia, dando la impresión de que los demás están ausentes al mismo tiempo. Pero sabemos por la neurociencia y los fundamentos de la psicología que nunca podemos observar y describir lingüístico-verbalmente de forma imparcial y "objetiva". Siempre adoptamos una perspectiva evaluativa, es decir, dadora de sentido – compulsiva (véase el capítulo 6.14.6 sobre la dialéctica de la toma de decisiones compulsiva y la obligación de justificar). Esto significa que sólo podemos hacer afirmaciones muy limitadas sobre el "nivel" en el que nos encontramos estamos, qué estamos *haciendo realmente*, etc. Pues en cuanto hacemos esto, estamos fuera del proceso original y proporcionamos metacognitivamente información sobre nuestra vida interior. El proceso real – cualquiera que haya sido – ya ha pasado en este punto. Esta es la aplicación de la primera frase del Dàodéjīng, o Tao Te King (véase cap. 2.5). Pero como sólo podemos hacer ciencia dentro de nosotros mismos, como sólo nuestras propias impresiones sensoriales son accesibles a nuestra conciencia, y como en última instancia vivimos y sentimos dentro de nosotros mismos, incluso en la comunicación con los demás – lo único que tenemos –, algunas cosas se quedan en el camino, ya que no hay dos personas que puedan compartir plenamente las mismas experiencias. La inducción, la abducción, la deducción y todos estos procesos tienen su importancia relativa a nivel analítico. En la realidad, sin embargo, probablemente se fusionan de forma perfecta e inseparable. Esto estudios neurocientíficos no cambian nada, porque actualmente no "miden" ni "reconstruyen" la calidad de la experiencia subjetiva en tiempo real, sino que sólo relacionan actividades cerebrales ruidosas entre sí. Las condiciones de estudio son muy limitadas en cuanto a su flexibilidad y no pueden compararse con la vida y las acciones cotidianas. Se trata (todavía) de un proceso muy difuso. Los procesos inductivo, abductivo y deductivo reflejan así (al menos) tres caras diferentes del mismo fenómeno, por lo que la abducción (véase el capítulo 2.3) casi nunca se menciona siquiera en la discusión estadística. La información y la puntuación también son sólo modelos o construcciones. Además, parece útil o incluso necesario ocuparse de otras formas de inferencia, como el *tetralema* y la *dialéctica*, para mirar más allá del propio horizonte, que a menudo es muy limitado en la ciencia, como demuestra no sólo Feyerabend (1976). Todo esto indica que la cognición ciertamente no tiene lugar sólo en los niveles de inducción, abducción y deducción, sino en un nivel superior integrador, para el que la ciencia todavía no ha acuñado un término adecuado y que sólo es parcialmente accesible a nuestra conciencia hasta cierto punto. Esto significaría en términos simplificados – leer al principio más Bateson y Hegel, y sólo más tarde leer a Popper y otros.

Así, en lugar de preguntar si hay algo subjetivo o algo objetivo, o una inducción o una deducción, se deberían aplicar los siguientes criterios ya mencionados y elaborados. Éstos no adoptan un papel particularmente típico ni para los bayesianos ni para los clásicos, ni para los cuantitativos frente a cualitativo, pero son criterios de calidad de la ciencia ampliamente relevantes y sobre ellos se puede leer en detalle en muchos libros de texto (entre otros, Steinke, 2000; Bergman & Coxon, 2005; Rocco, 2010; Anderson, 2017). Los criterios comunes son

- Flexibilidad y disposición a pensar y proceder de forma diferente y a cuestionarse a uno mismo
- Transparencia y responsabilidad
- Intersubjetividad, es decir, capacidad para debatir y disposición para el discurso
- Pertinencia práctica
- Discusión específica del contexto, es decir, especificación precisa de qué decisiones se toman en qué momento y con qué validez decisiones se toman y con qué pretensión de validez, ya que esto también puede variar dentro de una investigación. puede variar
- Comprobación de la realidad, es decir, aplicación de los datos a otros contextos o datos simulados
- Aprender del modelo, es decir, deducir algo nuevo a partir del estado actual de los conocimientos para (re)investigar el mismo tema. tema (de nuevo) y aumentar los conocimientos.
- Reproducibilidad y replicación

6.11 Integración de métodos y métodos mixtos

Un análisis bibliográfico sobre el tema de Bayes y los métodos mixtos sigue siendo escaso en la actualidad, como muestra fácilmente una breve búsqueda en Internet. una breve búsqueda en Internet mostrará fácilmente. En primer lugar, las explicaciones sobre estadística clásica (véase el capítulo 4) y su integración con los métodos mixtos, a saber, que una secuencia cuidadosamente planificada de estudios cualitativos y cuantitativos puede dar lugar a resultados muy diferentes. estudios cualitativos o cuantitativos puede sacar a la luz diferentes perspectivas de un objeto de investigación, de modo que los estudios (cualitativos) pueden sacar a la luz diferentes perspectivas de un objeto de investigación, de modo que la integración (dialéctica) de los resultados conduzca a una mayor ganancia de conocimientos (Erzberger y Kelley). (Erzberger y Kelle, 2003;Morse, 2003; Creswell y Clark, 2011, cap. 3 para una visión general de los posibles diseños). Los datos cualitativos también pueden considerarse cuantitativos y viceversa, si esto tiene sentido, está justificado en la pregunta de investigación y es metodológicamente factible (véase el capítulo 11.9 sobre la práctica metodológica del análisis secuencial para un posible caso de exclusión), para ello varios ejemplos de casos sobre AED y paradigma de codificación en la sección 5.5). En el contexto de la estadística bayesiana, existe también la elección de la(s) distribución(es) a priori, en la que la información cualitativa se "traduce" en hipótesis de distribución numérica. Esto se ha debatido varias veces desde distintas perspectivas (por ejemplo, véanse 6.3.2.3, 6.8.1.4, 6.12, 6.12.2) y se ha profundizado durante los debates sobre la subjetividad (véase 6.5.2; también 6.14.6). La integración de métodos en el contexto de Bayes puede diferenciarse según:

- Secuencia de diseños de investigación cuantitativos y cualitativos
- Examen de los datos de un paradigma con métodos del otro paradigma, si tiene sentido y es posible.
- Elección de la distribución a priori basada en información cualitativa.

El término "Integración Bayesiana de (datos) Cuantitativos y Cualitativos", o BIQQ (= Bayesian Integration of Quantitative and Qualitative (data)) para abreviar, se encuentra en la literatura (Humphreys & Jacobs, 2015). Los autores consideran que el teorema de Bayes básicamente ya funciona según los métodos mixtos, de modo que las estrategias de inferencia de ambos enfoques -- QUAN y QUAL – pueden describirse desde una perspectiva bayesiana. Por tanto, sólo se requiere un pequeño esfuerzo para integrar ambas estrategias

en un único marco. En particular, para las redes bayesianas (véanse los paquetes de R `deal` o `bnlearn`), el enfoque se recomienda para investigar la causalidad en contextos complejos (Pearl, 2009). Sin embargo, según Humphreys y Jacobs (2015, p.42)

> „Though conceptually simple, however, no integrated Bayesian approach like BIQQ has, to our knowledge, been developed or used for this purpose."

Aunque los autores consideran que el modelo BIQQ no puede salvar completamente todos los aspectos entre QUAN y QUAL, tiene el potencial de poder lograr la integración en cuatro niveles. Estos puntos (ibíd., p.43) pueden describirse de forma esquemática, y los autores se ocupan obviamente de las pruebas de causalidad:

1. La integración de los datos basados en la correlación y en el proceso hacia un modelo único de inferencia causal.
2. Los resultados inferidos de un tipo de datos pueden utilizarse para enriquecer sustancialmente las interpretaciones del otro tipo de datos.
3. Estudio de efectos (causales) y de estudios de muestras pequeñas, de modo que se mejoran considerablemente las afirmaciones sobre estimaciones de efectos medios para poblaciones. El nivel de casos se mantiene para las estimaciones de los efectos del tratamiento, de modo que las conclusiones siempre tengan una base empírica real.
4. Proporcionar directrices para diferentes diseños de estudios de métodos mixtos con el fin de optimizar el uso de los recursos.

Los autores siguen proporcionando código R reproducible para los estudios de casos. En conjunto, este enfoque parece prometedor y merece la pena ampliarlo, aunque por el momento sigue sin estar claro hasta qué punto se puede explicar realmente la causalidad "real" y qué contextos pueden (o no) resultar especialmente adecuados para este enfoque. La conversión de la información cualitativa en distribuciones y límites cuantitativos desempeña obviamente un papel importante en esto, casi uno de los asuntos centrales de los métodos mixtos. Esto incluye la elección de distribuciones a priori. Puesto que esto juega un papel tan relevante y ampliamente discutido en la estadística bayesiana, veamos el tema un poco más técnicamente y a lo largo de varias pequeñas líneas. En la Sección 6.15.2, veremos cómo el desarrollo puede basarse en el aprendizaje a partir de la experiencia.

6.12 Elección de distribuciones a priori

En la elección de una distribución a priori tiene lugar una transformación real de información contextual cualitativa en expresiones matemáticamente precisas. Cualitativamente hablando, es la reconstrucción del conocimiento experto o específico del dominio (O'Hagan, Buck, Daneshkhah, Eiser, Garthwaite, Jenkinson, Oakley, & Rakow, 2006) para integrar toda la información de tal forma que se pueda realizar una reducción razonable de la expectativa. Esto se hace en un punto en el tiempo antes de que haya datos disponibles. Se hacen suposiciones

- sobre el aspecto de una distribución,
- la ubicación o el punto de máxima densidad,
- la ponderación asociada de los rangos de valores y los
- rangos de exclusión (por ejemplo, establecer la Prior = cero)

Los supuestos distribucionales se formulan independientemente de si el conocimiento cualitativo se convierte en forma numérica o se utiliza un procedimiento objetivo guiado por reglas (Jaynes, 2003). Los supuestos distribucionales más detallados proporcionan información sobre la medida en que los valores extremos parecen plausibles, si una distribución es simétrica, de colas múltiples, etc. Por regla general, la información a priori es cada vez menos importante a medida que aumentan los datos empíricos, a menos que sea tan masiva y contraria a los datos empíricos que siga influyendo claramente en la información a posteriori incluso con muestras de mayor tamaño. En consecuencia, su influencia es muy alta, especialmente en el caso de información incompleta y baja, por lo que merece la pena invertir en la reconstrucción del conocimiento experto. Y esta situación debería corresponder a la regla, ya que la mayoría de los estudios trabajan con restricciones múltiples y las muestras no son infinitamente grandes. Del mismo modo, los estudios preliminares directos suelen ser escasos.

Para elegir una distribución a priori se siguen diversas estrategias. No existe una regla fija que *obligue a tomar* una distribución determinada para un tema cualitativo concreto. Esto es así aunque se disponga de muchos conocimientos sobre cómo se distribuyen determinados parámetros o se disponga de procedimientos normalizados, como los descritos para los factores de Bayes. Preferiblemente, se eligen *distribuciones a priori conjugadas*, si es posible, porque la distribución posterior resultante es conocida gracias al teorema de Bayes y, técnicamente hablando, su integral puede calcularse analíticamente. Ejemplos de distribuciones a priori conjugadas son la distribución beta, la distribución beta-binomial o la distribución normal (Wikipedia, 2019e con un resumen tabulado).

Los problemas reales suelen ser tan complejos que las distribuciones a priori conjugadas rara vez desempeñan un papel. En el caso de estas soluciones no analíticas, la distribución posterior se calcula numéricamente a través de las integrales para distribuciones a priori no conjugadas mediante simulación MCMC (véase el capítulo 6.13). Se trata de cadenas de Markov generadas mediante simulación Monte Carlo. Dada la potencia de cálculo de los ordenadores actuales, la elección de una distribución a priori conjugada ya no es necesaria, aunque puede ser elegante. Gracias a la simulación MCMC, la distribución a priori puede elegirse según las necesidades, lo que no debe equipararse a arbitrariedad. Antes de que existieran estas grandes capacidades informáticas, la estadística de Bayes se veía drásticamente limitada cuando una distribución posterior sólo podía calcularse mediante aproximación numérica a mano. Esto apenas es posible manualmente en la práctica. El uso generalizado de la estadística de Bayes en la actualidad se debe sin duda en gran medida a las capacidades informáticas disponibles hoy en día.

Las distribuciones a priori son modelos y deben utilizarse como tales. La elección de una distribución a priori debe hacerse siempre con bastante escepticismo (McElreath, 2015), y esto se aplica a todas las especificaciones. Por ejemplo, hay consecuencias si se elige una distribución a priori uniforme o no informativa. Así, desde una perspectiva teórica de la información, no existen distribuciones a priori que asignan la misma probabilidad a todos los valores posibles. En este caso, se puede actuar de forma más estúpida de lo que se es en realidad con la esperanza de acertar "objetivamente". Y entonces puede ocurrir que una distribución uniforme contenga rangos de valores y se considere probable en principio, pero que, desde un punto de vista técnico, no puede darse en absoluto. Debe tenerse en cuenta que un único valor de una distribución continua siempre recibe la probabilidad de cero, ya que existe un espectro infinito de datos continuos y, por tanto, el único dato recibe una probabilidad concreta de prácticamente cero. Por lo tanto, se eligen rangos con límites superior e inferior, para los que se puede determinar fácilmente una probabilidad como área bajo la curva. En efecto, una Prior que una toda la masa a cero es muy difícil, ya que significa que uno parece estar seguro de antemano con un 100% de certeza de que no hay ningún efecto y que ningún otro valor excepto cero parece posible. ¿Qué sentido tiene entonces la investigación? Me viene a la mente una cita de Benjamin Franklin (1706-1790), uno de los fundadores de los Estados Unidos de América y un verdadero multitalento, quien supuestamente comentó: "„In this world nothing can be said to be certain, except death and taxes“. El lector interesado puede plantearse tal pregunta de investigación y también cuáles son las consecuencias prácticas y si de una situación tan supuestamente segura se puede surgir algún beneficio sustantivo.

La elección de la Prior tiene implicaciones para el sobreajuste (s. cap. 6.8.3), como McElreath (2015, p.166) explica:

> „The root of the problem of overfitting is a model's tendency to get overexcited by the training sample. When the priors are flat or nearly flat, the machine interprets this to mean that every parameter value is equally plausible. As a result, the model returns a posterior that encodes as much of the training sample — as represented by the likelihood function — as possible."

Una variante para eludir esto es la *regularización*, es decir, limitar la Prior a áreas basadas en el contenido y plausibles. Por ejemplo, si se tiene la firme suposición de que un efecto apunta en una dirección determinada, se pueden excluir de antemano los efectos negativos. En ese caso, no quedan cubiertos por la Posterior, lo que puede tener un doble filo. Una regularización demasiado fuerte conduce a un ajuste insuficiente y posiblemente excluye áreas de datos relevantes que se producen empíricamente y, por lo tanto, pertenecen absolutamente al modelo. Por ejemplo, si se examinan los valores de CI, no se deja que la variable a priori se aplane sobre todos los valores posibles, sino que se limitan los valores de CI por encima de la mayor sobredotación y por debajo de la menor subdotación o discapacidad intelectual, ya que empíricamente no se dan valores fuera de estos rangos. Permitir valores de CI de 5 o incluso valores inferiores o superiores a 220 significa ahora actuar fuera del rango medible que puede examinarse con los tests de CI convencionales, que, por cierto, de todos modos sólo son insuficientes en los rangos superior e inferior. En otro caso, si se examina la estatura humana, en general se sabe cuáles son los límites inferior y superior para los adultos a partir de cierta edad: basta con mirar el Libro Guinness de los Récords. En él figuran las personas más altas y más bajas y sus medidas, actualizadas anualmente. Los valores fuera de estos rangos (en el caso de los adultos) pueden entonces excluirse completamente con un cierto límite de tolerancia y gran parte de la masa concentrada en un determinado rango. Así pues, antes de elegir una Prior plano, hay que considerar si se está permitiendo un rango de valores que en realidad no se encuentra empíricamente. En un caso – como el de la energía liberada por la explosión de una supernova nueva y desconocida – esto puede estar justificado, y en otros casos es completamente inútil y distorsiona los resultados del análisis incluso antes de que se recojan los datos. En el caso de la explosión de supernova, sin embargo, se podría establecer un límite inferior porque sin una cantidad mínima de energía (o masa de una estrella) no se produce ninguna supernova, pero dejar mucho margen para un nuevo límite superior si, por ejemplo, la masa estelar estimada así lo sugiere. Así pues, la combinación de distintas informaciones contextuales puede utilizarse para establecer una Prior. Esto también deja claro que la regularización *sólo* describe el proceso de creación de una Prior informada. Si este proceso falla, se producirá un sobreajuste o un subajuste extremo. Dependiendo del contexto, puede ocurrir que esté prohibido excluir categóricamente rangos de valores. Entonces se limita una distribución mediante el valor esperado y la desviación estándar o la forma general. El estrechamiento o ensanchamiento de una distribución va acompañado de una reducción o aumento de las probabilidades en los rangos de valores, de modo que los valores no pueden excluirse, pero pueden llegar a ser máximamente improbables a priori. El ejemplo anterior de la medición de una supernova – muy simplificado -- presumiblemente ya no utilizaría una distribución normal en los rangos extremos, sino que quizá simplemente se cortaría por abajo e introduciría una cierta asimetría por arriba para hacer justicia a las condiciones empíricas.

Por tanto, se puede limitar la amplitud de una distribución, como muestra la siguiente Figura 6.64 utilizando el ejemplo de la distribución normal, para introducir aquí un cierto escepticismo con respecto a los valores posibles y plausibles. Volvemos a los valores probados de CI con un valor esperado de $CI = 100$ y una desviación estándar de $\sigma = 10$. Las tres curvas de la Figura 6.64 sólo se diferencian en su desviación estándar de $\sigma = 10$, 5 y 2, respectivamente. Queda claro que, debido al cambio en la desviación estándar, la masa de la distribución se vuelve cada vez más concentrada en el valor esperado y, por tanto, los valores más extremos parecen menos plausibles. Es importante entender que no existe una única Prior universal correcta (McElreath, 2015, p.95). Las distribuciones a priori son una expresión del conocimiento contextual, resultan de estudios empíricos previos, surgen de suposiciones subjetivas o todo lo anterior. Proporcionan un punto de referencia desde el que ver los datos. Hay más de uno. Aunque obviamente hay distribuciones priores – véanse los argumentos anteriores – que no son muy apropiadas, hay otras tantas distribuciones a priori que son justificables y comparablemente apropiadas. Decidir "objetivamente" a partir de esta selección cuál es la mejor, llevería la relatividad fundamental de la verdad ad absurdum. Es mucho más importante

aprender de la elección de la Prior y de los datos recogidos para utilizar este conocimiento de forma fundamentada la próxima vez (ptII_quan_Bayes_regularizacion.r).

```
# different sigmas (normal distribution)
thetas <- seq(50,150,0.01)
dnorm1 <- dnorm(thetas, 100, 10)
dnorm2 <- dnorm(thetas, 100, 5)
dnorm3 <- dnorm(thetas, 100, 2)
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l")
plot(thetas, dnorm1, col="violetred3", type="l", ylim=c(0,0.22),
       pre.plot=grid(), bty="n", xlab=expression(mu), ylab="Density")
lines(thetas, dnorm2, lty=2, col="darkred")
lines(thetas, dnorm3, lty=3, col="blue")
legend("right", legend=c("10","5","2"),
       col=c("violetred3","darkred","blue"),
       lty=1:3, lwd=2,bty="n", horiz=FALSE, title=expression(paste(sigma)))
mtext(expression(paste("Different ",sigma,
       " of the normal distribution",sep="")), outer=TRUE, line=-2, cex=1.5, side=3)
```

Cuanto más amplia es la distribución, más valores extremos (simétricamente distribuidos) se admiten como plausibles y viceversa. En cambio, la Prior plana admite todos los valores por igual. No reconoce el escepticismo. En el caso extremo, sin embargo, podría elegirse una distribución normal tan amplia que fuera casi plana para el intervalo de valores que cabe esperar de forma realista, como puede verse en la Figura 6.65. Las tres curvas que allí se muestran son iguales. Las tres curvas que allí se muestran proceden de la misma distribución, una distribución normal N(0, 200). La curva superior muestra la forma de campana habitual y las dos curvas inferiores muestran sendas secciones, una para el intervalo de valores de -3 a +3 y otra para 200 a 203. Ambas parecen planas en este estrecho intervalo de 6 desviaciones típicas. *Allí* la distribución normal se comporta como una distribución uniforme. En todo el rango de valores, la curva vuelve a comportarse como una distribución normal. Las líneas discontinuas verticales en la curva central y superior muestran el mismo rango de valores de -3 a +3. Si una distribución se extiende de esta forma, el área alrededor del valor esperado sigue representando la mayor parte de la masa de la distribución, pero toda la densidad se distribuye en un área extremadamente grande. Esto debe justificarse y no debe aceptarse simplemente a priori como algo dado. Así, la densidad en los intervalos esperados empíricamente, por ejemplo de -2 a +2, es muy pequeña en comparación con una distribución normal con N(0, 1). De este modo, a los rangos de valores improbables se les asigna, posiblemente de forma innecesaria y artificial, una Prior excesiva que carece de base teórica y empírica (ptII_quan_Bayes_regularization.r).

```
# flat prior
thetas <- seq(-500,500,0.01)
dnorm1 <- dnorm(thetas, 0, 200)
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l", mfrow=c(3,1))
plot(thetas, dnorm1, col="violetred3", type="l", pre.plot=grid(),
       bty="n", xlab=expression(mu), ylab="Density")
abline(v=c(-3,3),col="blue",lty=2)
plot(thetas, dnorm1, col="violetred3", xli=c(-3,3), type="l",
       pre.plot=grid(), bty="n", xlab=expression(mu), ylab="Density")
abline(v=c(-3,3),col="blue",lty=2)
plot(thetas, dnorm1, col="violetred3", xli=c(200,203), type="l",
       pre.plot=grid(), bty="n", xlab=expression(mu), ylab="Density")
mtext(expression(paste("Flat Prior - various perspectives",sep="")),
       outer=TRUE, line=-2, cex=1.5, side=3)
```

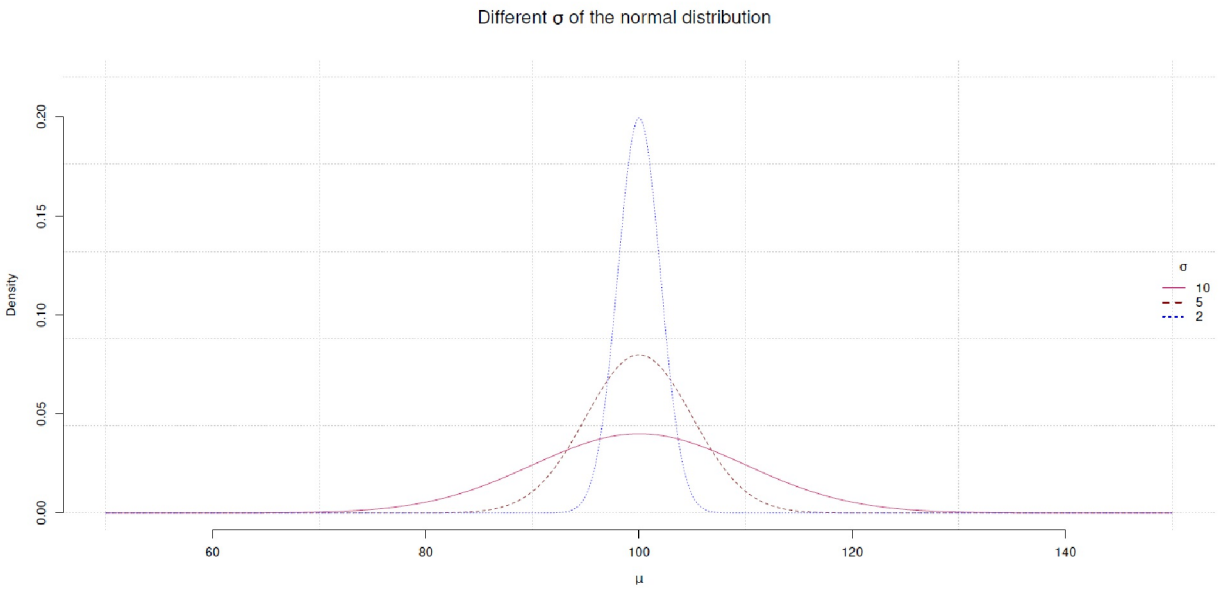


Figura 6.64. *Regularización (distribución normal, diferentes desviaciones estándar)*

Es una cuestión de perspectiva y limitaciones. Recuerde que una prioridad plana tiene prácticamente ningún efecto sobre la posterior. Este es el caso cuando las estimaciones frecuentistas se parecen a las de la estadística bayesiana. Una distribución normal de N(0, 1), por ejemplo, considera valores fuera de 2 ya improbables. Las funciones de distribución de estos cuantiles para desviaciones estándar de 1 a 6 ya son porcentajes de áreas muy escasas, como puede verse aquí – de odo unilateral y bilateral.

```
> # percent area below the curve for different N(0,1) sds from 1 to 4
> 2*(1-pnorm(1:6))
[1] 3.173105e-01 4.550026e-02 2.699796e-03 6.334248e-05
[5] 5.733031e-07 1.973175e-09
> 1-pnorm(1:6)
[1] 1.586553e-01 2.275013e-02 1.349898e-03 3.167124e-05
[5] 2.866516e-07 9.865877e-10
> pnorms <- sapply(1:2, function(i) i*(1-pnorm(1:6)))
> colnames(pnorms) <- c("one-sided","two-sided")
> pnorms
      one-sided    two-sided
[1,] 1.586553e-01 3.173105e-01
[2,] 2.275013e-02 4.550026e-02
[3,] 1.349898e-03 2.699796e-03
[4,] 3.167124e-05 6.334248e-05
[5,] 2.866516e-07 5.733031e-07
[6,] 9.865877e-10 1.973175e-09
```

A menudo se utilizan distribuciones muy amplias basadas en la distribución normal si la intención requiere una distribución uniforme, pero esto está fuera de lugar por varias razones y también se considera una Prior impropia. Esto significa que, sin limitar el rango de valores, una distribución uniforme no puede integrarse si va de $-\infty$ a $+\infty$. Sin embargo, todavía se puede calcular una Posterior en muchos casos. Como en el caso de la distribución uniforme, una Prior impropia corresponde a una Prior no informativa, mientras que una distribución normal muy amplia se considera una Prior débilmente informativa. Otro ejemplo de

una Prior impropia es el Beta(0; 0) o Prior de Haldane con valores extremos 0 y 1, que expresan ignorancia completa, es decir, cuando no está claro en absoluto si es posible el éxito o el fracaso y los valores de la distribución Beta son a → 0 y b → 0 respectivamente. La Prior de Haldane (Etz & Wagenmakers, 2017) corresponde a una distribución uniforme en la escala logit y fue aceptada como Prior útil por Jeffreys (Jeffreys, 1939/1961). Jaynes (1968) propuso la Prior Haldane como distribución a priori para expresar la incertidumbre completa sobre las probabilidades a priori, como cuando cuando, por ejemplo, no está claro si el éxito y el fracaso son siquiera posibles (Studer, 1996b).

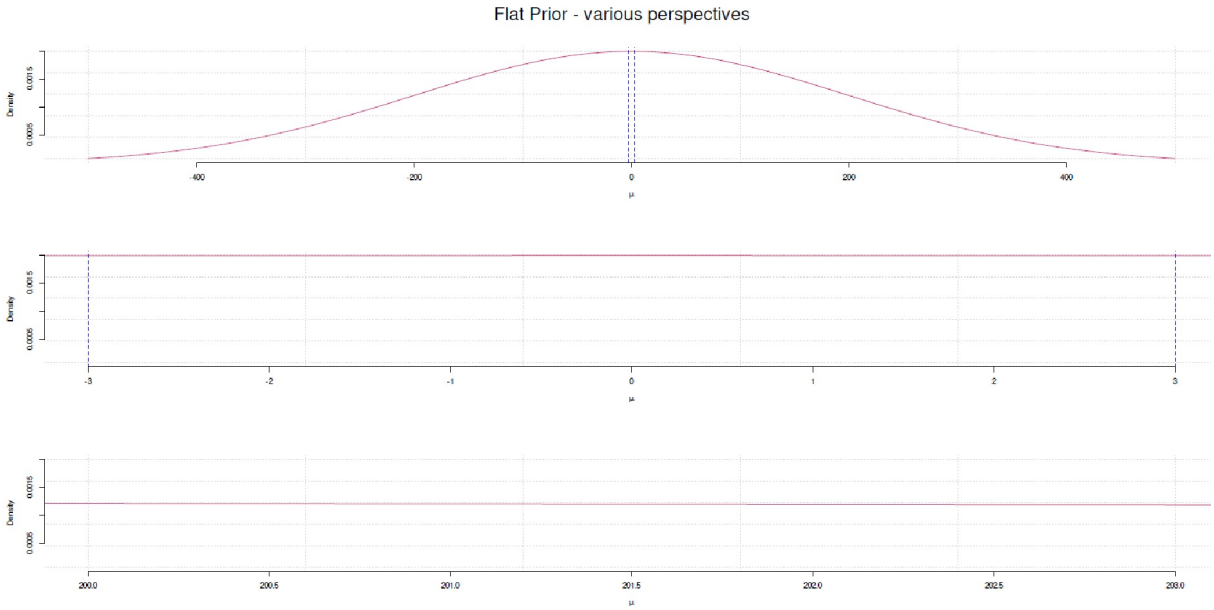


Figura 6.65. *Regularización (distribución normal como Prior plana)*

Si los rangos de valores válidos están limitados por la Prior con barreras (por encima, por debajo) y la Prior se elige por los parámetros de tal manera que sea cuasi plana, no hay ninguna diferencia real en el rango de valores de interés si se trata de una distribución uniforme (= distribución beta) o de una distribución normal o Cauchy extremadamente amplia y plana. Por otra parte, sólo la distribución uniforme es una Prior impropia. En función del coeficiente que se deba estimar se seleccionan distintas distribuciones a priori (por ejemplo, distribuciones normales para los pesos o distribuciones de Cauchy para la desviación estándar o los parámetros de varianza en modelos lineales, etc.). Como se puede ver en la Figura 6.66, la distribución normal *N*(0, 200), la Cauchy *Cauchy*(0, 200) y la uniforme tienen de nuevo un aspecto casi idéntico para el intervalo de valores de -3 a + 3, aunque en cada caso surgen diferencias entre las distribuciones y en ningún caso parecen idénticas.

```
# flat priors - different distributions
fak <- 1.3
thetas <- seq(-500,500,0.01)
dnorm1 <- dnorm(thetas, 0, 200)
dcauchy1 <- dcauchy(thetas, 0, 200)
dunif1 <- dunif(thetas,-500,500)
par(oma=c(2,1,2,1), "cex.axis"=1, bty="l", mfrow=c(3,2))
xlim <- c(-3,3)
ylim.dnorm <- c(0,max(dnorm1))
ylim.dcauchy <- c(0,max(dcauchy1))
ylim.dunif <- c(0,max(dunif1))
plot(thetas, dnorm1, col="violetred3", ylim=ylim.dnorm*fak, type="l",
       pre.plot=grid(), bty="n", xlab=expression(mu), ylab="Density")
plot(thetas, dnorm1, col="violetred3", ylim=ylim.dnorm*fak, xlim=xlim, type="l",
```

```
        pre.plot=grid(), bty="n", xlab=expression(mu), ylab="Density")
plot(thetas, dcauchy1, col="steelblue", ylim=ylim.dcauchy*fak,
        type="l", pre.plot=grid(), bty="n", xlab=expression(mu), ylab="Density")
plot(thetas, dcauchy1, col="steelblue", ylim=ylim.dcauchy*fak, xlim=xlim,
        type="l", pre.plot=grid(), bty="n", xlab=expression(mu), xlab="Density")
plot(thetas, dunif1, col="seagreen", ylim=ylim.dunif*fak, type="l",
        pre.plot=grid(), bty="n", xlab=expression(mu), ylab="Density")
plot(thetas, dunif1, col="seagreen", ylim=ylim.dunif*fak, xlim=xlim, type="l",
        pre.plot=grid(), bty="n", xlab=expression(mu), ylab="Density")
mtext(expression(paste("Flat Priors",sep="")), outer=TRUE,line=-1, cex=1.5, side=3)
mtext(expression(paste("Comparison of different distributions",sep="")),
        outer=TRUE,line=-3, cex=1, side=3)
```

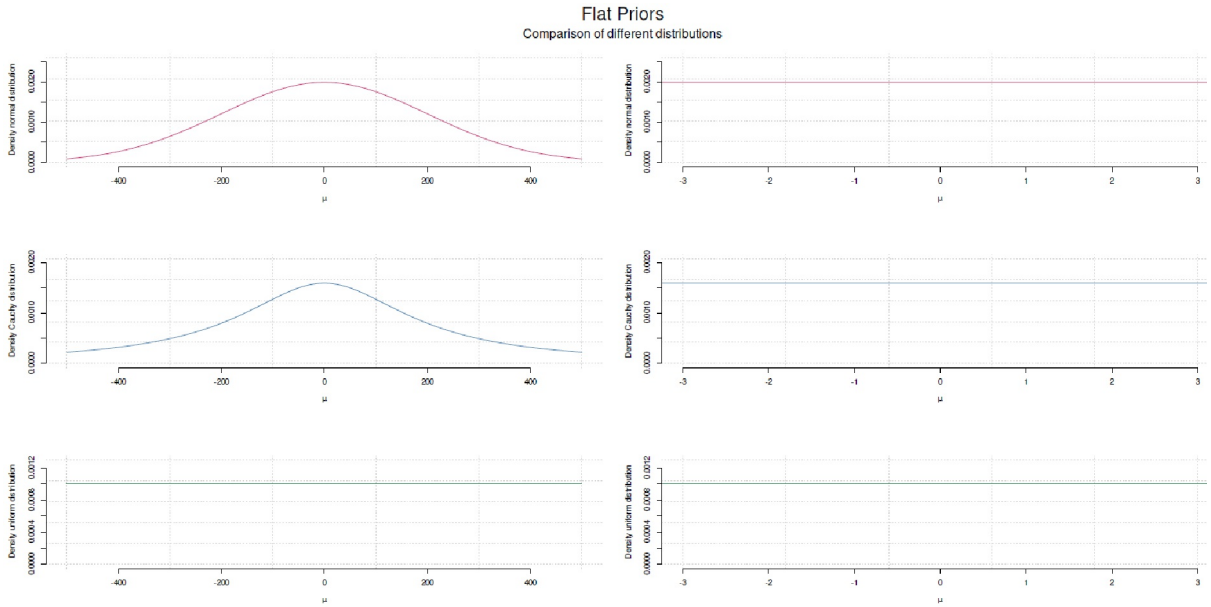


Figura 6.66. *Regularización (diferentes distribuciones como Prior plana)*

Gabry y Goodrich (2018-04-13) comentan a su vez la elección de distribuciones a priori en un artículo sobre modelización con `rstanarm`,

> „Rarely is it appropriate in any applied setting to use a prior that gives the same (or nearly the same) probability mass to values near zero as it gives values bigger than the age of the universe in nanoseconds. Even a much narrower prior than that, e.g., a normal distribution with $\sigma = 500$, will tend to put much more probability mass on unreasonable parameter values than reasonable ones. In fact, using the prior $\theta \sim \text{Normal}(\mu = 0; \sigma = 500)$ implies some strange prior beliefs."

Si se convierte esta cita en código R, se obtiene lo siguiente (véase la Fig. 6.67).

```
# no flat priors!
# http://mc-stan.org/rstanarm/articles/priors.html
lim <- 1500
theta <- seq(-lim,lim,0.1)
mu <- 0
sigma <- 500
norm.dens <- dnorm(theta, mean=mu, sd=sigma)
compare <- sigma/2
norm.tab <- data.frame(theta,norm.dens)
```

```
head(norm.tab)
tail(norm.tab)
# -lim < compare < lim
p <- 1 - 2*pnorm(-compare, mean=mu, sd=sigma)
p
colos <- adjustcolor(c("orange","darkred"),alpha=0.2)
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l")
plot(theta,norm.dens, type="l", col="darkred", bty="n",
       pre.plot=grid(), xlab=expression(paste(theta)),
       ylab="Density", cex.lab=1.2)
theta.outtake <- theta[norm.tab[,"theta"] >
       -compare & norm.tab[,"theta"] < compare]
dens.outtake <- norm.tab[norm.tab[,"theta"] >
       -compare & norm.tab[,"theta"] < compare,"norm.dens"]
theta.outt.l <- length(theta.outtake)
theta.startx <- theta.outtake[1]
theta.endx <- theta.outtake[theta.outt.l]
min.d <- min(norm.dens)
polygon(x=c(theta), y=c(norm.dens), col=colos[1], border=NA)
polygon(x=c(theta.startx, theta.outtake, theta.endx,theta.endx),
       y=c(min.d,dens.outtake,dens.outtake[theta.outt.l],min.d),
       col=colos[2], border=NA)
mtext("Prior distribution areas", outer=TRUE, line=-2,
       cex=1.5, side=3)
```

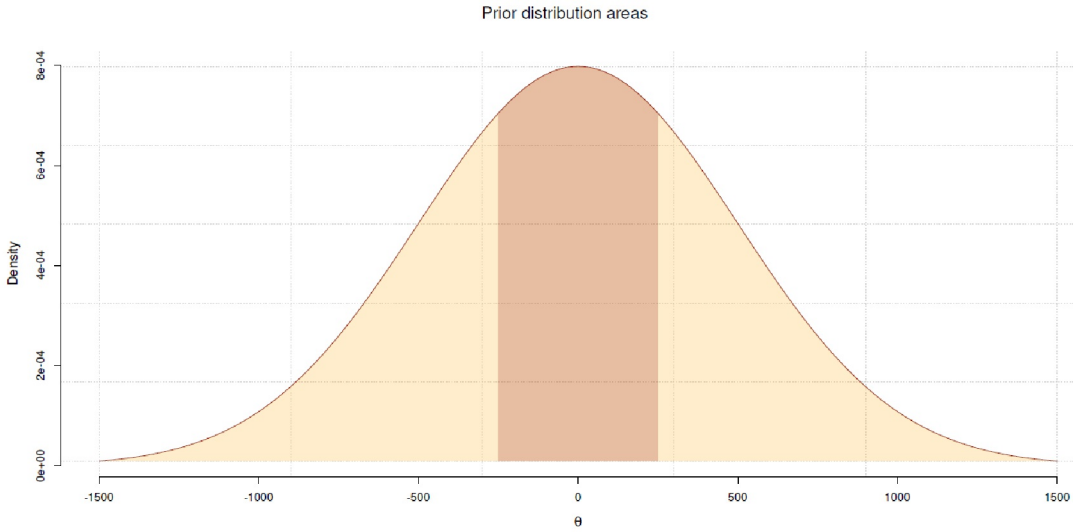


Figura 6.67. *Regularización (distribución normal amplia)*

Una distribución normal con $\mu = 0$ y $\sigma = 500$ ya es una distribución muy amplia comparada con una distribución normal estándar con $\sigma = 1$. Comparando esto con la Figura 6.67 anterior, debería quedar claro que se puede sustituir una distribución uniforme por una distribución muy amplia – como la distribución normal aquí – para ciertos rangos de valores. Sin embargo, que esto tenga sentido es otra historia. Si tomamos la ley de conservación de la energía, que establece ante todo que hay una cantidad finita de recursos y que debemos utilizarlos con prudencia, podemos plantearnos en qué estamos invirtiendo a la hora de formular una Prior. ¿Tiene sentido incluir datos no plausibles o excluir rangos de valores con una cierta límite de tolerancia? ¿Realmente se desea que la curva sea completamente plana, o se prefiere invertir tiempo en la cuestión de qué información cualitativa contextual conduce a qué conclusiones a priori y cómo pueden aplicarse luego estas conclusiones como hipótesis de distribución? Una consecuencia sería que uno se hace vulnerable a los ataques porque no pretende no saber nada, sino que se atreve a comprometerse con un valor o rango de valores. Sin embargo, esto es lo que posibilita el principio del examen crítico, de la falsación. Una curva plana que no prefiere ni excluye nada no puede falsificarse. Es prácticamente siempre o nunca cierta. En algunos ejemplos de este libro utilizamos una distribución uniforme, pero sobre todo cuando no tenemos

ni idea de un tema y no es didácticamente una cuestión de los propias Priors en este punto. Entonces la distribución uniforme no es necesariamente plausible, pero puede implementarse fácilmente para demostrar otra cosa. Cuando, con un poco de tiempo invertido, es posible obtener al menos una visión básica de los rangos de valores, prescindimos inmediatamente de una distribución uniforme o increíblemente amplia, que a su vez es plana en el rango de valores de interés. En consecuencia, Gabry y Goodrich (2018-04-13) dan una recomendación sobre la magnitud al formular una Prior, "una heurística es establecer la escala un orden de magnitud mayor de lo que se sospecha que es – y tiene el beneficio añadido de ayudar a estabilizar los cálculos." Una discusión discusión de la influencia de las Priors débilmente informadas en las inferencias es proporcionada por R-code Betancourt (2017-01).

Si existe la suposición de una distribución bastante amplia y, por lo tanto, la inclusión plausible de valores más extremos debido a una muestra de entrenamiento, esto debería cuestionarse críticamente y confrontarse con consideraciones teóricas. Sabemos (no sólo) por la estadística frecuentista lo mucho que las muestras pueden diferir entre sí y, por lo tanto, existe el peligro de sobreajuste (véase el capítulo 6.8.3). Una recomendación habitual es restringir la amplitud de la distribución a priori y regularizarla para adoptar un punto de vista conservador. Este punto de vista conservador consistiría en que los valores de la Prior se acercaran entre sí y se concentraran en torno a aproximadamente cero como representantes del máximo escepticismo. El efecto cero, como sabemos, no se da en la práctica. Esto refuerza la hipótesis de la localización, ya que en ella la densidad se hace mayor. Si se exagera, hablamos de subajuste (véase el capítulo 6.8.3), es decir, que se incluye muy poca información contextual como en una distribución uniforme que no prefiere nada en absoluto. Del mismo modo, en vista de la información disponible, no tiene mucho sentido suponer cero como representante de los parámetros si hay indicios de lo contrario desde el punto de vista contextual. Es diferente suponer un efecto pequeño o ningún efecto en absoluto. No queremos repetir los errores del NHST (véase el capítulo 4.3.8). Por ello, contraponemos al escepticismo razonable una cierta disposición a asumir riesgos y a aceptar la posibilidad del fracaso y, por tanto, de la falsificación. El reto consiste en integrar y equilibrar el escepticismo y el riesgo. Así lo formula McElreath (2015, p.187, énfasis en el original),

> „Regularizing priors are great, because they reduce over tting. But if they are too skeptical, they prevent the model from learning from the data [y esto lleva al subajuste, nota de los autores]. So to use them most effectively, you need some way to tune them. Turning them isn't always easy. If you have enough data, you can split into 'train' and 'test' samples and then try different priors and select the one that provides the smallest deviance on the test sample. That is the essence of cross-validation, a common technique for reducing overfitting."

Dado que, según el teorema de Bayes, la distribución posterior es proporcional al producto de la probabilidad y la probabilidad a priori, la introducción de una distribución uniforme (Prior), por ejemplo, conduce a una distribución posterior *completamente* basada en los datos: la Likelihood domina por completo el cuadro. *Esto corresponde al caso de la estadística frecuentista*, en la que ninguna información contextual ponderada puede entrar en las ecuaciones. La información contextual entra cuando se ponderan ciertos datos de forma diferente. Sin embargo, si sólo hay unos pocos datos o los datos proceden de una muestra mal elegida, este caso sería una situación desfavorable y no daría una Prior adecuada. En este caso, habría que invertir en establecer una distribución a priori plausible traduciendo la información cualitativa en suposiciones numéricas (véase el capítulo 6.14.6 para ejemplos que abordan la relación entre Prior, Likelihood y Posterior).

Así pues, el compromiso entre el exceso y la falta de ajuste consiste en combinar información realista (Prior), por un lado, con los datos empíricos (Likelihood) y sus características, por otro. De este modo, el modelo global puede aprender de los datos, pero se limita a un rango plausible (Prior). A medida que crece el conjunto de datos, estas restricciones tienden por sí solas en una dirección empíricamente sólida – a medida que se va disponiendo de nueva información, la Prior puede, por supuesto, reformularse y sus límites pueden redefinirse.

Además, existe el requisito de planificar la elección de las distribuciones a priori desde una perspectiva informativa, una perspectiva muy defendida por Jaynes (1957a, 1957b). Así queda claro que una distribución

a priori siempre contiene información y que no puede haber distribuciones no informativas. No se trata de trabajar bayesianamente por un lado y pretender por otro lado, como si hubiera que minimizar la influencia de la Prior, porque perturba, posiblemente afecte subjetivamente, etc. Todo es información y, en consecuencia, todo es una decisión más o menos consciente – por ejemplo, que se fije o no un determinado enfoque, que la elección recaiga sobre una clase de procedimientos para el análisis de datos, que se acote tal o cual rango de valores, que una Prior tenga esta forma, etc. Esto significa que la densidad máxima se concentra en uno o varios puntos o rangos de valores concretos, y este rango puede ser cero, aunque no necesariamente. Hay muchas preguntas que se centran en valores específicos por lo que el valor cero resulta muy rápidamente irrelevante y poco informativo. Lo mismo ocurre con las distribuciones a priori planas.

6.12.1 La distribución beta

Para tener una mejor idea de los supuestos de distribución y sus implicaciones, se mostrará para los dos parámetros a y b que determinan la distribución beta. La distribución beta es una distribución extremadamente flexible que puede adoptar formas muy diferentes (véase la Fig. 6.68). Éstas van desde la Haldane-Prior-Beta(0, 0), la forma cercana de Haldane con Beta(0.001, 0.001), la Beta(1, 1) uniforme hasta la Prior-Beta(0.1, 0.1) Jeffrey, pasando por formas de distribución asimétricas, simétricas o incluso abiertas hacia un lado. Esto permite representar una gran cantidad de conocimiento previo si la elección de a y b se hace con cuidado (`ptII_quan_Bayes_Beta-distribution.r`).

```
# plot various beta distributions with varying a, base
xaxis <- seq(0,1,length.out=1000)
a <- c(0.001,0.01,0.1,1,5,seq(0.5,5,by=0.25),100)
b <- c(0.001,0.01,0.1,1,5,seq(5,0.5,by=-0.25),100)
ab <- data.frame(a,b)
ab
ab.dim <- dim(ab)
colo <- rainbow(ab.dim[1])
areadim <- ceiling(sqrt(ab.dim[1]))
par(mfrow=c(areadim, areadim), oma=c(1,1,5,1), cex.axis=0.8)
for(i in 1:ab.dim[1])
{
plot(xaxis, dbeta(xaxis, ab[i,"a"], ab[i,"b"]), type="l",
        col="violetred3", xlab="", ylab="", bty="n",
        main=paste("a = ",ab[i,"a"],", b = ",ab[i,"b"],sep=""),
        pre.plot=grid())
}
mtext(expression(paste("The various shapes of Beta-distributions",
        sep="")),
        3, line=1.6, cex=1.5, outer=TRUE)
```

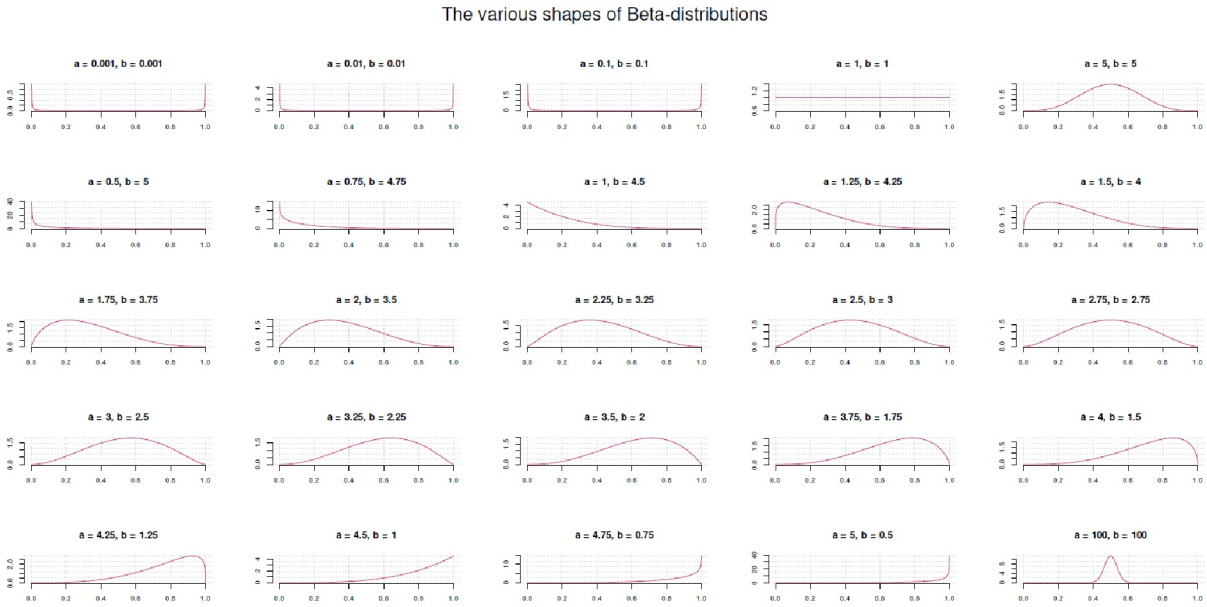


Figura 6.68. *Distribuciones Beta*

La fórmula de la distribución Beta es similar a la de la distribución binomial, pero se complementa con la función Beta, que a su vez consiste en un cociente de producto de funciones gamma (véase la Fig. 6.69, `ptII_quan_Bayes_Gamma-distribution.r`, código R no impreso). La función Beta o integral euleriana del primer género describe una función de dos números complejos, cada uno de los cuales debe tener una parte real positiva. Su definición es

$$B(a,\, b) = \frac{\Gamma(a) \cdot \Gamma(b)}{\Gamma(a+b)} \tag{6.84}$$

$$= \int_0^1 t^{a-1} \cdot (1-t)^{b-1}\, \mathrm{d}t \tag{6.85}$$

La densidad de probabilidad de la distribución Beta Beta(a, b) se define como

$$f(x)_{\text{Beta}} = \begin{cases} \dfrac{1}{B(a,\, b)} \cdot x^{a-1} \cdot (1-x)^{b-1} & 0 \le x \le 1;\, a,\, b > 0 \\ 0 & x < 0 \text{ oder x} > 1 \end{cases} \tag{6.86}$$

El valor esperado es

$$\mathbb{E}(X)_{\text{Beta}} = \frac{a}{a+b} \tag{6.87}$$

y la varianza

$$\text{Var}(X)_{\text{Beta}} = \frac{a \cdot b}{(a+b+1) \cdot (a+b)^2} \tag{6.88}$$

La función gamma Γ(x) o integral de Euler del segundo tipo no está definida uniformemente. Como integral impropia para números complejos x con parte real positiva, su denominación es la siguiente definición

$$\Gamma(x) = \int_0^\infty t^{x-1} \cdot e^{-t}\,\mathrm{d}t \tag{6.89}$$

Las funciones gamma incompletas no regularizadas son con límite superior

$$\gamma(a,\, x) = \int_0^x t^{x-1} \cdot e^{-t}\,\mathrm{d}t \tag{6.90}$$

y con límite inferior

$$\Gamma(a,\, x) = \int_x^\infty t^{x-1} \cdot e^{-t}\,\mathrm{d}t \tag{6.91}$$

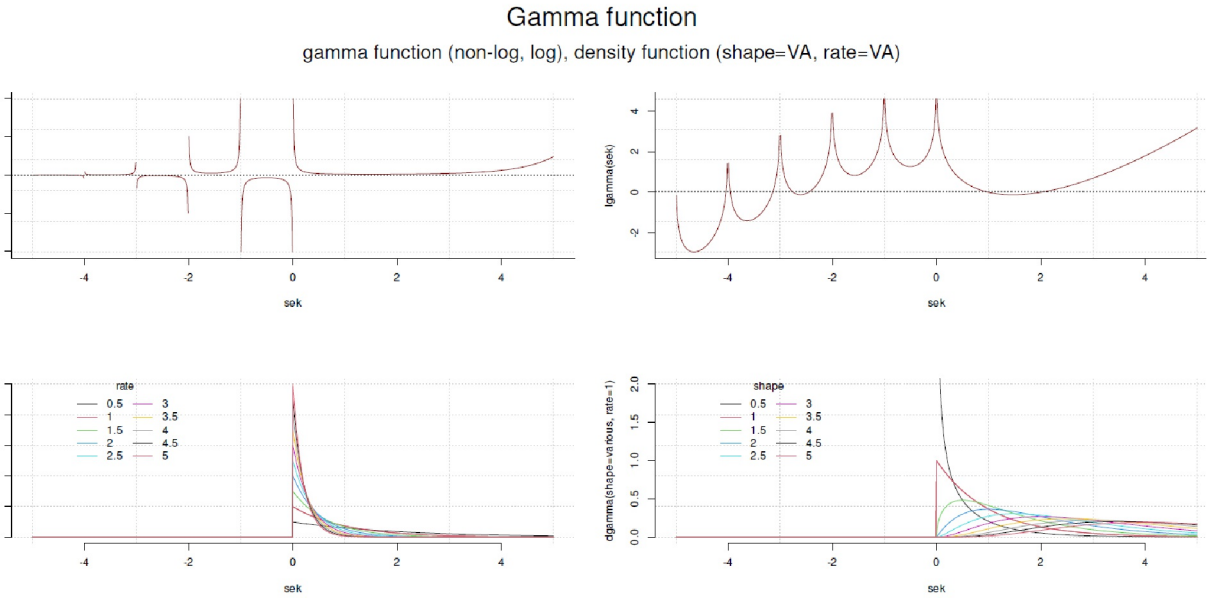


Figura 6.69. *Distribución gamma (función gamma así como función de densidad, varios parámetros para shape y rate)*

La densidad de probabilidad de la distribución gamma Γ(p, b) es

$$f(x)_{\text{Gamma}} = \begin{cases} \dfrac{b^p}{\Gamma(p)} \cdot x^{p-1} \cdot e^{-bx} & x,\, p,\, b > 0 \\ 0 & x \leq 0 \end{cases} \tag{6.92}$$

con el valor esperado

$$\mathbb{E}(X)_{\text{Gamma}} = \frac{p}{b} \tag{6.93}$$

y varianza

$$\mathrm{Var}(X)_{\mathrm{Gamma}} = \frac{p}{b^2} \tag{6.95}$$

Para comparar, veamos la distribución binomial discreta con tamaño *n* y probabilidad *p* (= éxito)

$$B(k \mid p,\ n) = \begin{cases} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} & \text{para k } \epsilon \{0, 1, ..., n\} \\ 0 & \text{si no} \end{cases} \tag{6.96}$$

con el valor esperado

$$\mathbb{E}(X)_{\mathrm{Binomial}} = n \cdot p \tag{6.97}$$

y varianza

$$\mathrm{Var}(X)_{\mathrm{Binomial}} = n \cdot p \cdot q \tag{6.98}$$

$$= n \cdot p \cdot (1-p) \tag{6.99}$$

ya que

$$q = 1 - p \tag{6.100}$$

Como alternativa a la distribución Beta, también podrían utilizarse distribuciones triangulares y rectangulares o, en principio, cualquier distribución que se pueda justificar razonablemente. La elección de la distribución está relacionada con el modelo de generación de datos que se supone eficaz. Se formulan supuestos diferentes para los procesos de descomposición en biología que para las pruebas en psicología o los cursos de procesos económicos y escenarios políticos.

La correlación – solución analítica – de la distribución Beta y la distribución binomial puede mostrarse en el proceso de actualización de la Prior por la Likelihood a la Posterior. Como se ha mencionado, la distribución Beta es una Prior conjugada en el contexto de una Likelihood binomial. Se trata de una conjugación si la Posterior pertenece al mismo tipo de distribución que la Prior. Cuando se combina una Prior distribuida de modo Beta y datos binomiales (= Likelihood) este es el caso para obtener una Posterior Beta a partir de una Prior Beta (Etz, 2015b, con código R).

$$\mathrm{Prior}_{\mathrm{Beta}} = p^{a-1} \cdot (1-p)^{b-1} \tag{6.101}$$

$$\mathrm{Likelihood}_{\mathrm{Binomial}} = p^{s} \cdot (1-p)^{n-s} \tag{6.102}$$

Ya que

$$\mathrm{Posterior} \propto \mathrm{Prior} \cdot \mathrm{Likelihood} \tag{6.103}$$

se puede combinar las formulas para Prior y Likelihood como

$$\mathrm{Posterior}_{\mathrm{Beta}} = p^{a-1} \cdot (1-p)^{b-1} \cdot p^{s}(1-p)^{n-s} \tag{6.104}$$

$$= p^{a-1} \cdot p^{s} \cdot (1-p)^{b-1} \cdot (1-p)^{n-s} \tag{6.105}$$

El resultado es la Posterior, que se obtiene directamente de los parámetros *a* y *b* de la Prior y *x* y *n* de la Likelihood. Esto puede escribirse en una función y aplicarse fácilmente.

$$\text{Posterior}_{\text{Beta}} = p^{a-1+s} \cdot (1-p)^{b-1+n-s} \tag{6.106}$$

Lo elegante es que a la Likelihood no le importa el orden en que se utilizan los datos. Esta característica se transfiere de la Likelihood a la Posterior. Por lo tanto se puede calcular el mismo resultado final a partir de todos los datos y la Prior que si los datos se reciben según el proceso de actualización anterior (véase también el ejemplo de los datos sobre las tasas de aprobados del centro de terapia `start again`, Cap. 6.8.4.4 o 6.15.2). Si una Prior tiene la forma Beta(a, b) y el proceso binomial tiene *s* aciertos y *n* ensayos, la Posterior resulta en a Beta(a + s; b + n - s) -- y ya está. Volvamos a Lady Bristol (véanse los capítulos 4.3.2.1, 6.3.2.5 y 6.13.4). Si, por ejemplo, consigue s = 8 aciertos con n = 8 ensayos, sólo necesitamos los parámetros *a* y *b* de la distribución Beta para determinar la Prior y, posteriormente, directamente la Posterior para Lady Bristol. Así que existen fórmulas para calcular a partir de media y la varianza de una distribución Beta. También existen fórmulas para determinar *a* y *b* a partir de valores *s* y *n* especialmente formulados del modelo binomial, para expresar una expectativa a priori reordenando las fórmulas en cada caso. (Ejemplos con código R: `CrossValidated`, usuario Dave Kincaid, 2011). Veamos esto en R (`ptII_quan_Bayes_Fisher_LadyBristol-Distribución beta.r`):

```
# Lady Bristol - choose a prior
# we believe the median of the prior is 0.85
quantile1 <- list(p=0.5, qua=0.65)
# we believe the 99.999th percentile of the prior is 0.95
quantile2 <- list(p=0.8, qua=0.8)
# we believe the 0.001st percentile of the prior is 0.60
quantile3 <- list(p=0.2, qua=0.25)
prior.ab <- beta.determine.opt(p=c(0.5,0.8,0.2), qua=c(0.65,0.8,0.25),
            ab.start=c(1,1), graph=TRUE)
prior.ab
```

Como Prior elegimos $p = 0.65$ para la mediana y los límites $p = 0.8$ para el cuantil del 80% y $p = 0.25$ para el cuantil del 20%. La distribución resultante mediante la función R `beta.determine.opt()` descrita a continuación es una beta(1:241; 0:952). A partir de los valores s = 8 y n = 8 se obtiene a = 9 y *b* = 1 con `bino.ab.lik()` (véase la Fig. 6.70).

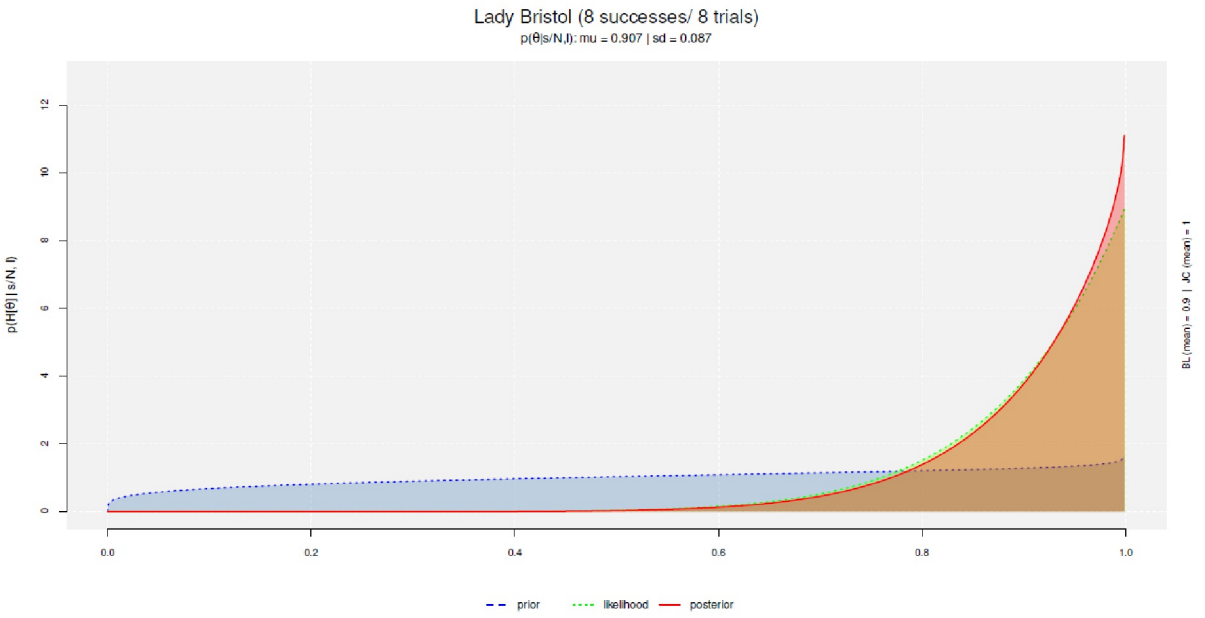


Figura 6.70. *Experimento del té de Lady Bristol*
(8 aciertos de 8 ensayos con Prior, Likelihood y Posterior).

```
# Lady Bristol - use empirical data for likelihood
si <- 8
Ni <- 8
lik.ab <- bino.ab.lik(si=si, Ni=Ni)
```

La Beta(9, 1) resultante es también una distribución abierta a la derecha y describe la Likelihood. El resultado Posterior es una Beta(9.241, 0.952) con `bino.ab.post()`. La figura 6.70 muestra las tres distribuciones utilizando `beta.triplot()`.

```
# post
post.ab <- bino.ab.post(a.prior=prior.ab$res.ab3["a"],
        b.prior=prior.ab$res.ab3["b"], si=si, Ni=Ni)
```

De este modo se obtienen los valores para la Priori, la Likelihood y la Posterior. El código R para el gráfico (véase la Fig. 6.70) está disponible en el script R.

```
> # output
> prior.ab$res.ab3
a b
1.2414975 0.9520068
> lik.ab
a b
9 1
attr(,"type")
[1] "likelihood"
> post.ab
a b
9.2414975 0.9520068
attr(,"type")
[1] "post"
```

y un resumen de la Posterior

```
> # summary
> post.summary <- beta.summary(a=post.ab[["a"]], b=post.ab[["b"]])
> unlist(post.summary)
a           b          mode       mean        sd         var
9.24149746 0.95200676 1.00585747 0.90660653 0.08697304 0.00756431
```

¿De qué otra forma se pueden obtener los valores de *a* y *b* si no se tienen en cuenta los porcentajes de éxito? Una posibilidad es aproximar *a* y *b* a partir de la mediana supuesta y dos cuantiles para los límites superior e inferior. Coghlan (2017, cap. 2.2) proporciona una función de este tipo `findBeta()` basada en `beta.select()` del paquete `LearnBayes` de R. Este ejemplo nos hace conscientes de que merece la pena invertir en la eficacia de la programación en R. Reproducimos los valores del autor y medimos el tiempo que tarda el ordenador en encontrar los valores de *a* y *b* (véase la Fig. 6.71). Para ello utilizamos el R (`ptII_quan_Bayes_find-Beta-distribution-shapeparams.r`):

```
> # call original code from website - slow...
> # we believe the median of the prior is 0.85
> quantile1 <- list(p=0.5, x=0.85)
> # we believe the 99.999th percentile of the prior is 0.95
> quantile2 <- list(p=0.99999,x=0.95)
> # we believe the 0.001st percentile of the prior is 0.60
> quantile3 <- list(p=0.00001,x=0.60)
> system.time( a.b.values <- findBeta(quantile1,quantile2,quantile3) )
[1] "The best beta prior has a= 52.22 b= 9.52105105105105"
```

```
User System verstrichen
8.168 0.000 8.167
```

Ahora tomamos una versión beta.determine(), que hemos mejorado y que funciona exactamente igual que `findBeta()`, pero trata los valores vectorizados y no funciona con bucles (véase también Ligges, 2005).

```
> # call and measure time
> system.time(a.b.values <- beta.determine(p=c(0.5,0.99999,0.00001),
        + qua=c(0.85,0.95,0.60)))
User    System      verstrichen
3.260   0.005       3.261
> a.b.values
        a       b
3001 52.22 9.521051
```

Ahora se añade una tercera versión llamada `beta.determine.opt()`, que utiliza la función de optimización `optim()` como componente central. En caso de que se quiera, la función genera un gráfico (véase la Fig. 6.71).

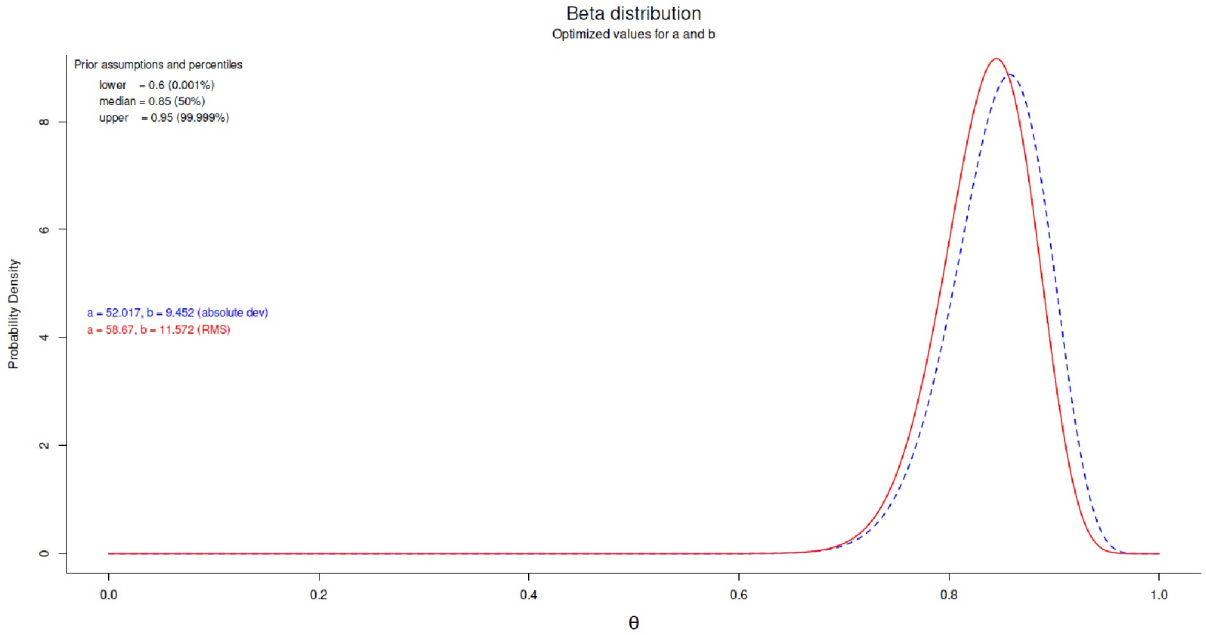


Figura 6.71. *Distribución Beta (Parámetros a y b a partir de los cuantiles y la mediana)*

```
> # call and measure time
> system.time(a.b.values.opt <- beta.determine.opt(
        + p=c(0.5,0.99999,0.00001), qua=c(0.85,0.95,0.60),
        + ab.start=NULL, graph=TRUE) )
User    System      verstrichen
0.048   0.000       0.048
> a.b.values.opt
$quans
    p       x
1 0.50000 0.85
2 0.99999 0.95
3 0.00001 0.60

$res.ab
    a       b
52.016712 9.451692
```

```
$res.ab3
    a       b
58.66955 11.57242

$quans.prior
                a      b
q.1.2.prior 110.81 19.83
q.1.3.prior 52.22 9.49

$res.optim
$res.optim$par
    a       b
52.016712 9.451692

$res.optim$value
[1] 0.02418102
$res.optim$counts
function gradient
201 NA
$res.optim$convergence
[1] 0
$res.optim$message
NULL
```

Como se puede ver, los valores difieren ligeramente. Los tiempos del sistema, sin embargo, difieren drásticamente; gracias a la vectorización sigue requiriendo un ≈39.93% y gracias a `optim()` sólo un 0.59% del tiempo de cálculo original, un aumento de velocidad por un factor de 2.5 a 170. Los valores difieren de los aquí reportados dependiendo del sistema informático, pero el aumento de velocidad debería ser comparable.

```
> 3.261/8.167
[1] 0.3992898
> 0.048/3.261
[1] 0.01471941
> 0.048/8.167
[1] 0.005877311
```

El paquete de R `rriskDistributions` incluye varias funciones que permiten determinar los parámetros del modelo de funciones de interés según este principio. Un vistazo al código fuente muestra que todas ellas utilizan `optim()` y permiten determinar distribuciones utilizando más de dos cuantiles.

```
get.beta.par(p=c(0.00001,0.5,0.99999), q=c(0.60,0.85,0.95))
```

La figura 6.72 muestra que estas estimaciones corresponden a las expectativas teóricas (Rcode no impreso). Aquí se han representado los valores de `beta.determine.opt()` y de `get.beta.par()` junto con la curva teórica de la distribución Beta. El gráfico de la izquierda contiene esto mientras que el gráfico de la derecha ha distorsionado ligeramente los valores estimados con `jitter()` para ver las diferencias más claramente. Como se puede ver, esto es difícilmente posible. El método método funciona muy bien.

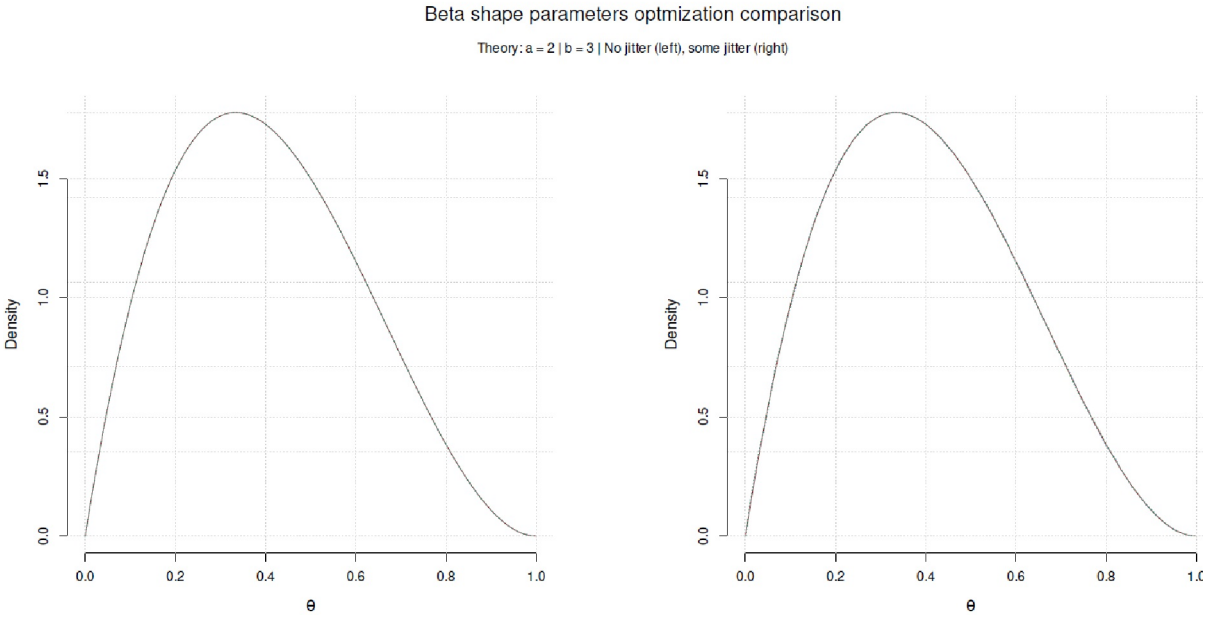


Figura 6.72. *Distribución beta (comparación con las estimaciones de* `beta.determine.opt()` *y* `get.beta.par()`)

6.12.2 Relación Prior-Likelihood-Posterior

Ya hemos sentado las bases de la relación entre las tres distribuciones. El objetivo debe ser proporcionar al proceso general de análisis información válida procedente de una amplia variedad de fuentes. La Figura 6.73 muestra cómo puede ser esto (véase también la Fig. 6.134). El siguiente ejemplo de datos examina el no éxito frente éxito con los datos ficticios de 10 éxitos en 14 ensayos. Elegimos una distribución Beta para la Prior y el proceso binomial para la Posterior. Al igual que la Prior, la Posterior es conjugada para la Likelihood, ambas se originan a partir de la distribución Beta. Los valores de la Posterior tomamos de la salida de `bino.abs()`. El código R está contenido en el archivo de guión `ptII_quan_Bayes_Beta-distribucion.r`:

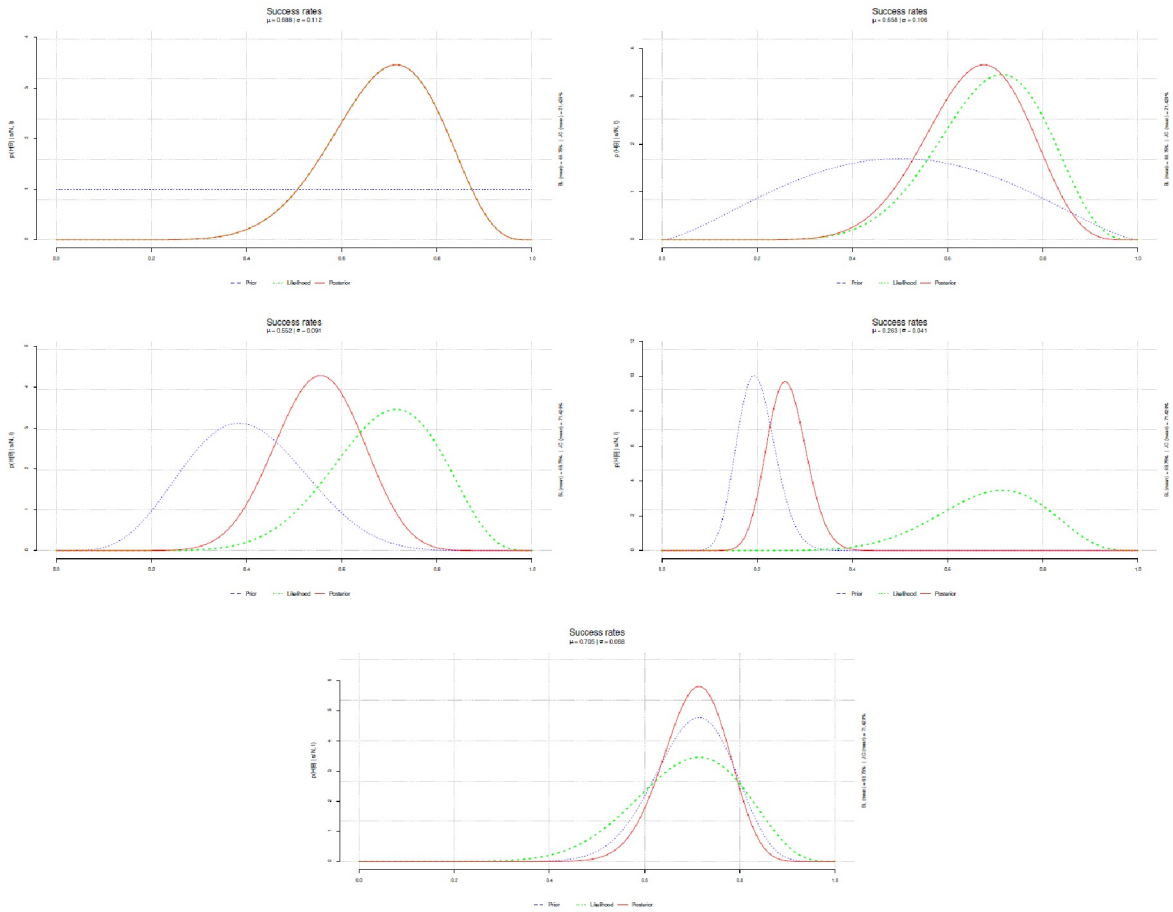


Figura 6.73. *Regularización (Influencia de la Prior)*

```
# influence of priors
# example success rates
# empirical data
si <- 10
Ni <- 14
prior.vs <- data.frame(theta.prior=c(0.5,0.5,0.4,0.2,0.7),
       nprior=c(2,5,15,100,30))
prior.vs
# 1- uniform prior
# 2- less heavily informed prior, different than likelihood
# 3- heavily informed prior, different than likelihood
# 4- extreme informed prior, different than likelihood
# 5- heavily informed prior, same as likelihood
par(ask=TRUE)
for(i in 1:dim(prior.vs)[1])
{
 bino.abs.res <- bino.abs(si=si, Ni=Ni,
       theta.prior=prior.vs[i,"theta.prior"],
       nprior=prior.vs[i,"nprior"], graph=FALSE)
 v <- bino.abs.res$res
 beta.triplot(si, Ni, v, filling=FALSE)
}
```

Se puede interpretar la figura 6.73 de forma que, dependiendo de la Priori, la Likelihood tiene una influencia mayor o menor. Está claro que la Posterior es un compromiso ponderado entre la Prior y la probabilidad a posteriori. Enumeramos algunas posibilidades y damos pistas sobre cómo evaluar la situación respectiva:

- Una distribución uniforme (arriba a la izquierda) asigna la misma probabilidad a priori a todos los valores. Por tanto, la Prior domina por completo la Posterior. La influencia de la Priori es despreciable, ya que no especifica ningún lugar. Este caso sólo debe elegirse si todos los valores potenciales tienen realmente la misma probabilidad. Dado que este caso corresponde a la maximización de la Likelihood, aquí también se podría trabajar estadísticamente de forma clásica. Las estimaciones son las mismas. No se cumple el propósito de la Prior, a saber, formar una expectativa razonable a priori. De hecho, se ha una Posterior, pero sin ninguna ponderación sustantiva por parte de la Prior.
- Una Prior ligeramente informada (arriba a la derecha) favorece aquí un porcentaje de éxito del 50 %, sin excluir especialmente los demás valores ni considerarlos muy improbables. Los valores marginales reciben cada uno una probabilidad reducida. La probabilidad puede determinar la Posterior en consecuencia, pero sin darle forma completamente como en el caso uniforme. Queda por ver si esto tiene sentido en el caso real. Una probabilidad del 50 % con dos resultados posibles (éxito, fracaso) corresponde al lanzamiento de una moneda al aire. El contenido de una Prior tan amplia debería prepararse mejor para formular una expectativa más clara basada en el contenido.
- Una Prior fuertemente informada con suposiciones que difieren de la probabilidad (centro izquierda) favorece el 40% y estrecha el rango considerablemente (centro izquierda), excluyendo así casi por completo los rangos de datos. Como resultado, la Posterior se acerca notablemente a la Prior y la Likelihood tiene menos influencia. Este sería el caso si el conocimiento previo apuntara en una dirección diferente que los datos recogidos. Este sería el caso, por ejemplo, si las variaciones aleatorias en experimentos repetidos conducen a resultados diferentes y los modelos aún no pueden caracterizarse como estables. También puede ocurrir que la Likelihood contenga un valor atípico o que la Prior se base en suposiciones falsas y que la Likelihood ofrezca en realidad el mejor modelo. En este punto habría que investigar más a fondo la hipótesis corto-empírica, ya que no es necesariamente posible asumir un resultado final. También sería necesario investigar cómo de grandes pueden ser las fluctuaciones entre colecciones de datos para comprender mejor el caso de los valores atípicos frente a los límites de adecuación del modelo.
- Una prior extremadamente informada con supuestos diferentes de la probabilidad (en el centro a la derecha) favorece al 20% y proporciona una cantidad sustancial de información como base. Así, el rango es muy estrecho y una gran parte de los valores posibles se marca como prácticamente improbable. A consecuencia es que la Posterior se inclina mucho hacia la Prior y la Likelihood ejerce poca influencia. Este sería el caso si se dispusiera de muchos datos y conocimientos a priori, y aquí la Likelihood es más bien una especie de valor atípico hacia arriba. Esto influye, pero da forma a la Posterior sólo de forma limitada. Esta situación sería una continuación de la situación anterior de la Prior fuertemente informada, cuando se dispone de más datos y las estimaciones se estabilizan.
- Una prior fuertemente informada con supuestos cuasi idénticos a los de la Likelihood (abajo) favorece al 70%, de modo que Prior, Likelihood y Posterior están cuasi superpuestos. La Prior apoya la Likelihood, de modo que la Posterior es aún más evidente y, en consecuencia, tiene una varianza menor y tanto la expectativa general como los datos empíricos apuntan en una misma dirección. Este sería el caso si, dado un conocimiento previo suficiente, una réplica experimental arrojara exactamente los mismos resultados, lo que hablaría de efectos extremadamente estables. Un ejemplo sería el rendimiento de la memoria del cerebro en la tradición de la investigación de Hermann Ebbinghaus (1850-1983).

La influencia de la Prior se puede observar en el capítulo 6.15.2 en un estudio de caso empírico -las tasas de éxito de la institución start again- a lo largo de más de 20 años. Allí se puede ver (véase la Fig. 6.134,) cómo, con la estabilización de las estimaciones, las Priores y las Posteriores difieren sólo ligeramente entre sí. A lo largo de los años, la Likelihood muestra valores diferentes cada año, pero básicamente estas fluctuaciones ya no conducen a un cambio sustancial en las estimaciones, incluso con valores atípicos ocasionales hacia arriba o hacia abajo. Básicamente, o bien las condiciones económico-sociales, la clientela, las características estructurales de la organización o similares tendrían que cambiar en una medida exorbitante para desplazar la Posterior significativamente hacia arriba o hacia abajo, o bien tendrían que producirse cambios que se desvíen de la Posterior actual durante un periodo de tiempo muy largo. Esto no es imposible, pero requiere una justificación sustancial. Cuanto más se acerquen la Prior, la Likelihood y la

Posterior más convergen las desviaciones a través del modelo (sobreajuste, infraajuste). Una distribución a priori uniforme puede ser ciertamente una forma de underfitting en relación con un modelo estable a posteriori, al menos en relación con la Prior. Los demás casos descritos, sin embargo, no pertenecen al overfitting, ya que no estamos examinando la adecuación del modelo, sino sólo la cuestión de la relación entre Prior, Likelihood y Posterior. Los modelos elegidos son modelos apropiados para la cuestión de las probabilidades de éxito.

Ahora ampliamos nuestra visión y nos centramos sólo en la sustancia de la Prior, es decir, en la cantidad de información que entra en ella. Así, la estimación de un experto puede constituir una predicción sólida si contiene mucha información. Si sólo se dispone de una única investigación y su calidad es controvertida, se puede considerar la Prior poco sólida. Si se dispone de una rama de la investigación bien documentada con muchos estudios cuyos resultados apuntan todos más o menos en la misma dirección, la Prior contiene una gran cantidad de información. Para simplificar nos quedamos en el terreno de las suertes binomiales (tasas de éxito) y las distribuciones Beta conjugada. Se combinan tres casos diferentes de tasas de éxito (= Likelihoods) con distribuciones a priori de distinto tamaño. El parámetro de la Prior se elige $\theta = 0.5$. Aumentamos el tamaño de la Prior de 2 a 10, 100, 500 y 5 000. Las curvas tienen el siguiente significado: azul (Prior), verde (Likelihood) y rojo (Posterior). El primer caso supone cero aciertos en 2 ensayos ($si = 0$, $Ni = 2$). En este caso, la Prior domina casi por completo a la Posterior. La Likelihood no tiene prácticamente nada que decir (`ptII_quan_Bayes_Prior-Likeli-Post_relationship.r`), como ilustra la Figura 6.74.

```
# with "delay" of the influence
# of the likelihood on the posterior
# depends on the N
npriors <- c(2,10,20,100,500,5000)
# use this to show how the prior dominates the posterior
si <- 0
Ni <- 2
par(oma=c(2,1,4,1), "cex.axis"=1, bty="l", mfrow=c(3,2))
for(i in npriors) bino.abs(si=si, Ni=Ni, theta.prior=0.5, nprior=i,
        rn=paste("nprior = ",i,sep=""), graph=TRUE)
mtext("Various Priors and resulting Likelihood + Posterior",
        outer=TRUE, line=1.7, cex=1.3, side=3)
mtext(paste(paste("si = ",si," | N = ",Ni," | priors = [",sep=""),
        paste(npriors, collapse=", "),"]",sep=""),
        outer=TRUE, line=-0.4, cex=1, side=3)
```

El segundo caso supone 7 aciertos en 23 ensayos (si = 7, Ni = 23). En este caso tarda bastante tiempo hasta que la Prior describa la Posterior casi por completo. Al principio, es sobre todo Likelihood (véase la Fig. 6.75).

```
# use this to show how the prior dominates more and more the likelihood
# the likelihood dominates at the beginning
si <- 7
Ni <- 23
par(oma=c(2,1,4,1), "cex.axis"=1, bty="l", mfrow=c(3,2))
for(i in npriors) bino.abs(si=si, Ni=Ni, theta.prior=0.5, nprior=i,
        rn=paste("nprior = ",i,sep=""), graph=TRUE)
mtext("Various Priors and resulting Likelihood + Posterior",
        outer=TRUE, line=1.7, cex=1.3, side=3)
mtext(paste(paste("si = ",si," | N = ",Ni," | priors = [",sep=""),
        paste(npriors, collapse=", "),"]",sep=""),
        outer=TRUE, line=-0.4, cex=1, side=3)
```

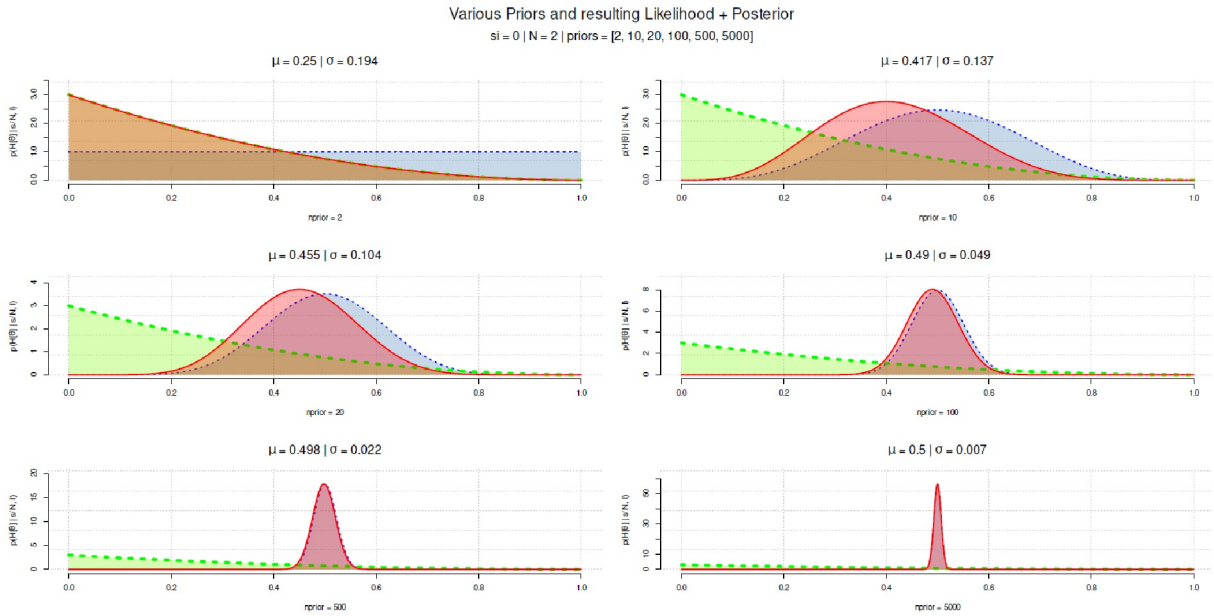


Figura 6.74. *Prior, Likelihood y Posterior (0 aciertos, 2 ensayos)*

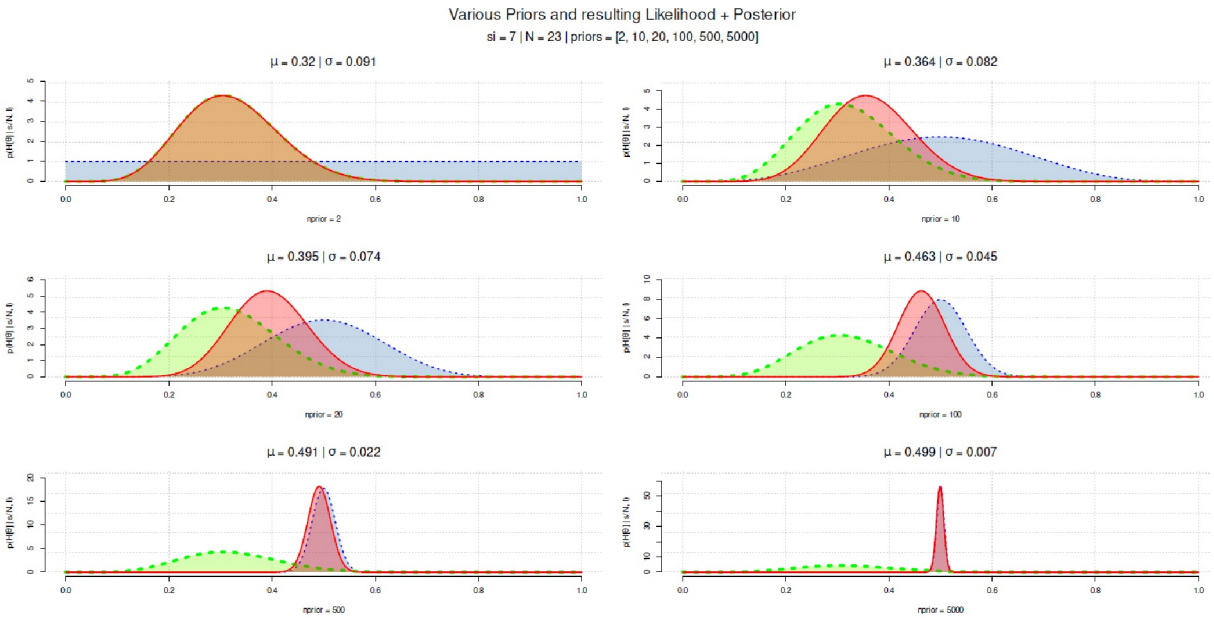


Figura 6.75. *Prior, Likelihood y Posterior (7 aciertos, 23 ensayos)*

El tercer caso supone 700 aciertos en 1 000 ensayos (si = 700, Ni = 1000). En extremación del caso anterior, tarda aún más tiempo hasta que la Prior se impone (véase la Fig. 6.76).

```
# use this to show how the likelihood dominates the posterior
# and slowly the prior becomes influential
si <- 700
Ni <- 1000
par(oma=c(2,1,4,1), "cex.axis"=1, bty="l", mfrow=c(3,2))
for(i in npriors) bino.abs(si=si, Ni=Ni, theta.prior=0.5, nprior=i,
       rn=paste("nprior = ",i,sep=""), graph=TRUE)
mtext("Various Priors and resulting Likelihood + Posterior",
       outer=TRUE, line=1.7, cex=1.3, side=3)
mtext(paste(paste("si = ",si," | N = ",Ni," | priors = [",sep=""),
       paste(npriors, collapse=", "),"]",sep=""),
       outer=TRUE, line=-0.4, cex=1, side=3)
```

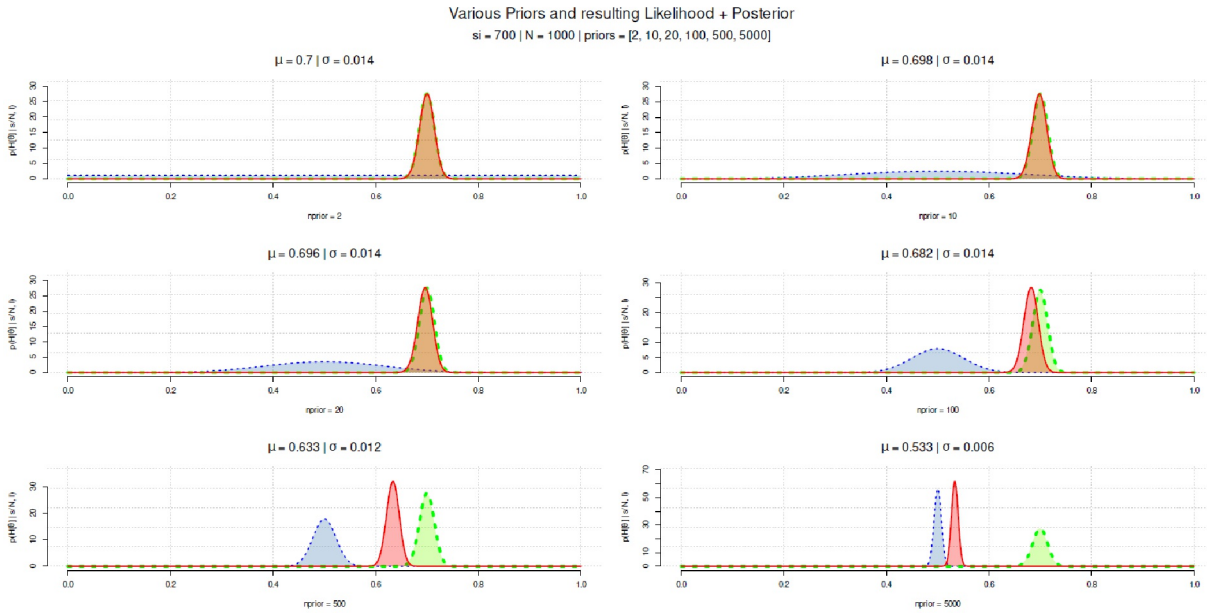


Figura 6.76. *Prior, Likelihood y Posterior (700 aciertos, 1000 ensayos)*

> Tarea 6.9: Comportamiento de la Posterior
>
> La tarea para el lector consiste ahora en comprender los cambios gráfiko por gráfico con el trasfondo de las explicaciones anteriores y explicar las afirmaciones anteriores sobre *por qué* la Posterior se comporta de la forma en que lo hace.

Ahora bien, se podría suponer que el tamaño de la muestra es suficiente para que la influencia de la Prior sobre la Posterior se desvanezca en favor de la Likelihood, ya que para el parámetro θ de interés, así como para el conjunto de datos $\mathbf{X}$, se mantiene

$$p(\theta \mid \mathbf{X}) \propto p(\mathbf{X} \mid \theta) \cdot p(\theta) \tag{6.107}$$

y por lo tanto en la escala `log()`

$$\log(p(\theta \mid \mathbf{X})) = c + \log(p(\mathbf{X} \mid \theta)) + \log(p(\theta)) \qquad (6.108)$$
$$= c + \mathcal{L}(\theta; \mathbf{X}) + \log(p(\theta))$$

La Likelihood $L(\theta , X)$ aumenta con el tamaño de la muestra como una función directa de los datos, mientras que la información a priori no lo hace, sino que permanece constante. A medida que aumenta el tamaño de la muestra, aumenta el valor absoluto de $L(\theta , X)$, mientras que para un valor θ fijo la Prior no cambia. Así, la influencia de la Likelihood crece en detrimento de la Prior.

Como señala incansablemente Jaynes (2003), la estadística bayesiana no trata sólo de cálculos cuantitativos, sino de información. Una Prior uniforme puede verse influida inmediatamente por la Likelihood. En cambio, una Prior informativa puede crear una imagen completamente distinta e implica regularización. La propia Prior contiene información no basada en los datos empíricos, que no cambia a pesar de los cambios drásticos por parte de la Likelihood y, por lo tanto, es eficaz incluso con muestras grandes. Ejemplos pueden ser Priores mal elegidas, por ejemplo, si favorecen un rango que no es realísticamente posible, como "personas de más de 3 m", o incluso no mensurable, como "valores de CI superiores a 250", etc., y así introducen permanentemente información perturbadora en el teorema de Bayes, que no es fácil de equilibrar con los datos existentes. También sería concebible que la información previa simplemente excluye áreas de datos realistas, lo que conduce a una incompatibilidad de los datos y las suposiciones. Una demostración sencilla de cada caso se ofrece en un debate sobre CrossValidated

(por el usuario Tim, 2016). Aquí se necesita entonces un tamaño de muestra mucho mayor en comparación con una Prior sabiamente elegida para producir una Posterior realista.

Así que, en general, la apariencia de la Posterior tiene que ver tanto con el tamaño de la muestra como con la información previa y los datos empíricos. Este es un comportamiento deseable de la Posterior, porque es la única forma de que la información previa pueda seguir siendo eficaz a pesar de tamaños de muestra mayores, aunque contradiga a la muestra mayor. Esta última podría estar sujeta a errores y posiblemente (por ejemplo, debido a la selección de la muestra) excluir valores legítimos. Por tanto, no se trata de un defecto sino, al contrario, de una ventaja del estadístico de Bayes. Imaginemos que empíricamente sólo se obtienen datos que representan un abanico bastante estrecho de posibilidades, por ejemplo porque la selección de la muestra es desfavorable, como se acaba de señalar. Una Prior informativa que explícitamente no sólo permite sino que posiblemente incluso favorece datos fuera de los datos empíricos, crea un equilibrio, una memoria y, por tanto, una incongruencia que espera explicación. Si las muestras grandes – que, dependiendo de su selección, aún pueden estar sujetas a error y no cubren todo el espectro de posibilidades – fueran suficientes por sí solas, uno podría ahorrarse la multiplicación por la Prior en el teorema de Bayes. Pero ocurre lo contrario. La Prior equilibra los datos para llegar a una Posterior razonable y significativa. Por eso es tan importante elegir una Prior cuidadosamente y con gran deliberación. También queda claro que una Prior uniforme es a menudo una mala elección porque no representa "no saber", sino posiblemente una actitud de "no me importa". En ese caso, la Prior uniforme no sólo deja que la Likelihood cree completamente la Posterior, sino algo mucho peor: da a cada rango de datos la misma probabilidad.

Pero cualquiera que trabaje empíricamente sabe que esto no es así. ¿Dónde, aparte de en un experimento con una moneda ideal, un dado ideal, una baraja ideal, etc., se da que todas las expresiones tengan realmente la misma probabilidad de manifestarse y que ésta pueda incluso extenderse sobre un espectro cuasi infinito? Lo que hace (entonces) la Prior es "meramente" tomar una posición en cuanto a qué parte del espectro de datos se considera probable. Si se piensa más en esta figura, inmediatamente queda claro por qué el trabajo no se detiene con la Posterior, sino por qué deben (¡deben!) seguir las comprobaciones predictivas posteriores (posterior predictive checks, véase el capítulo 6.8.4.3) para comprobar la razonabilidad de los resultados encontrados sobre la base de los datos y no sólo mediante una réplica dirigida y los intentos de falsificación del propio modelo.

Todo el asunto puede complicarse arbitrariamente examinando la robustez de la Posterior dados distintos Priores (y parámetros del modelo) cuidadosamente elegidos. Por ejemplo, se podría comparar un modelo más conservador con un modelo progresivo (Spiegelhalter, 2004). Sin embargo, no se trata de saber qué Priores son superiores y cuál se toma, lo que sería una elección desfavorable, porque no se trata de elegir una Prior que a uno le guste, sino una que refleje fielmente la realidad y el conocimiento existente. En el sentido de

los modelos complejos (véase la discusión de los tipos de error en este contexto, sección 4.3.3.2), se debería crear más bien un intervalo de tolerancia para poder tratar la incertidumbre existente de forma selectiva y comparar diferentes escenarios (por ejemplo, el caso ideal, el peor caso, etc.) entre sí con respecto a los resultados y las medidas que se deriven.

Hay una última razón por la que debe elegirse una Prior: los datos previos, los conocimientos previos y la combinación de ambos. No importa si se trate de datos de otros estudios que se puedan introducir en la Prior o de cuidadosos supuestos teóricos y derivaciones, la Prior garantiza que el trabajo no se base ciegamente en los datos, sino que haya una continuidad entre el conocimiento previo y la posterior renovación y actualización de los conocimientos generales. El objetivo es un proceso de *aprendizaje acumulativo* en el que cada paso posterior se basa en el anterior. Y esto es exactamente lo que permite la combinación de la Prior y la Likelihood basada en nuevos datos, sin tirar por la borda los conocimientos previos y las ideas recopiladas con tanto esfuerzo.

En consecuencia, la elección de las Priores no debe hacerse por razones matemáticas, sino por razones de contenido. En una detallada entrada de blog de Lindeløv (2018), el autor utiliza código R para mostrar diferentes variantes para calcular los factores de Bayes. Al hacerlo, discute la elección de Priores para los diversos parámetros del modelo. Por ejemplo `get_prior()` del paquete `brms` de R muestra los valores elegidos automáticamente por `brm()`, que, sin embargo, no están informados por defecto. Ahora bien, encontrar los valores adecuados no es tarea fácil, como señala Lindeløv (ibíd.):

> „These priors do not at all represent our knowledge about patients, treatments, and the design. There's a big literature on howto set priors, and I'm not toowell acquainted with it. For a truly cumulative science, you would probably try to conduct meta-analyses on all published data and use those parameter estimates as your prior. This would be 'maximally informative priors' (as opposed to 'uninformative priors'). In practice,many set a prior using the results from one highly similar study or just make an even vaguer "expert judgment" to save time. Personally, to save time on smaller projects, I look at the 50 % and 95 % credible intervals and run it by a few colleagues to get a consensus summary of the current knowledge."

De este modo, el autor cubre tanto los requisitos teóricos, es decir, los metaanálisis detallados, como las realidades prácticas de la investigación cotidiana, a saber, el discurso con los colegas sobre los valores previos y el uso de estudios modelo comparables. Llegados a este punto, nos gustaría destacar la contribución del análisis cualitativo de datos a la hora de encontrar una Prior, especialmente cuando la información es mínima o está muy distribuida. El arsenal de métodos correspondiente se describe detalladamente en la Parte III. Cabe destacar el análisis de secuencias de la hermenéutica objetiva (s. cap. 11) y la necesidad de comprobar su contribución a la reconstrucción cualitativa precisa de la información previa.

Para pasar de la Prior, los datos y la Likelihood a la Posterior, necesitamos ahora el denominador del teorema de Bayes, la evidencia o verosimilitud total. En el caso de los modelos más complejos, esto es prácticamente imposible de conseguir analíticamente, por lo que se utilizan como estándar simulaciones MCMC.

6.13 Simulaciones Markov Chain Monte Carlo – MCMC

Los algoritmos MCMC proporcionan un amplio arsenal de métodos para aproximar numéricamente una Posterior y definir el espacio de parámetros de la forma más exhaustiva posible. De este modo, la Posterior se crea a partir de una muestra suficientemente grande, aleatoria y representativa (Kruschke, 2015b, p.143). Esto es necesario porque el denominador normalizador del teorema de Bayes a menudo ya *no* puede resolverse analíticamente en modelos complejos. En ese caso, el denominador sólo puede aproximarse mediante integración numérica y, a través de ella, determinarse la Posterior. Así, en lugar de encontrar una

solución exacta (por ejemplo, en Bretthorst, 1993, sobre la prueba *t* bayesiana exacta, o Pham-Gia, Turkkan y Eng, 1993, sobre la prueba de proporción bayesiana exacta), basta con una aproximación numérica suficientemente exacta. Esto requiere muchas muestras para trazar el espacio de parámetros y es muy intensivo desde el punto de vista informático. MCMC tiene cadenas de Markov de dos elementos y simulación Monte Carlo.

Las *cadenas de Markov*, en honor al matemático ruso Andrej AndreyevicMarkov (1856-1922), describen un modelo estocástico en el que la secuencia condicional de sucesos probables para cada suceso depende de los estados del suceso o sucesos actuales o anteriores. Así, los estados pasados pueden no influir en el estado actual, o pueden influir en él o hacerlo de una determinada manera. En general, los sucesos anteriores no deben condicionar el suceso actual, ya que es la única forma de generar muestras independientes. Por lo tanto, pueden surgir problemas si aparecen inesperadamente autocorrelaciones en muchos sucesos y, de este modo, los sucesos dependen unos de otros, aunque esto no es deseable. Las transiciones de los estados dentro de las cadenas de Markov se almacenan en *matrices de transición* o *gráficos de transición* (véase la Fig. 6.77F).

Estas matrices denotan las probabilidades de pasar de un suceso *i* al siguiente suceso *j*. Un ejemplo simple es el lanzamiento de una moneda con la probabilidad de transición $p = 0.5$ para llegar del estado *t* al estado $t + 1$ cuando se considera en el tiempo – o del estado *i* al estado *j* o permanecer en el estado *i* cuando se diferencia por estados. En el caso discreto, los sucesos pueden distinguirse entre cara y cruz. Una moneda justa dice poco sobre si el siguiente lanzamiento también es cara si ya lo era en el lanzamiento actual. En ese caso, los lanzamientos individuales no dependen unos de otros. Otro ejemplo clásico es el paseo aleatorio (= random walk). En este caso se camina un paso (= estado) a la izquierda (= i + 1) o a la derecha (= i - 1) con probabilidad *p* o $q = 1 - p$. Este principio básico puede complicarse a voluntad, tanto para el caso discreto como para el caso continuo, con varias variables y por tanto dimensiones, etc.

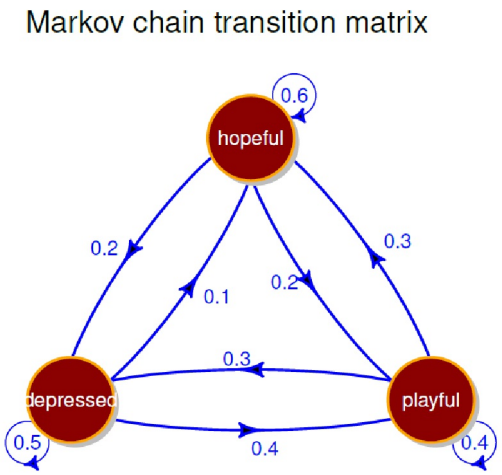


Figura 6.77. *Cadena de Markov – Gráfico de transición*

Se puede simular fácilmente un simple paseo aleatorio (`ptII_quan_Bayes_RandomWalk.r`), como muestra Figura 6.78:

```
> # simple random walk
> walk.2d <- randomwalk(seed=667)
> walk.3d <- randomwalk(seed=766, D=3)
> head(walk.3d)
       no x  y  z
[1,]    1  0  0  0
[2,]    2 -3  4 -3
[3,]    3  3  7  5
[4,]    4 13 12  0
```

```
[5,]    5  5  4  8
[6,]    6 -2 12  6
> tail(walk.3d)
           no    x   y   z
[995,]    995   7  -6  72
[996,]    996  12 -15  73
[997,]    997   5 -12  83
[998,]    998  12 -19  76
[999,]    999   5 -24  78
[1000,]  1000 -1 -23  71
```

Las simulaciones de MonteCarlo se basan en un gran número de experimentos aleatorios, cada uno de los cuales se basa en el mismo principio de cómo se producen los acontecimientos. La base fundamental es, como siempre, la ley de los grandes números. Por ejemplo, es posible aproximar numéricamente el número π de esta manera (Eddelbüttel, 2012) o resolver las integrales necesarias para la estadística de Bayes numéricamente en lugar de analíticamente.

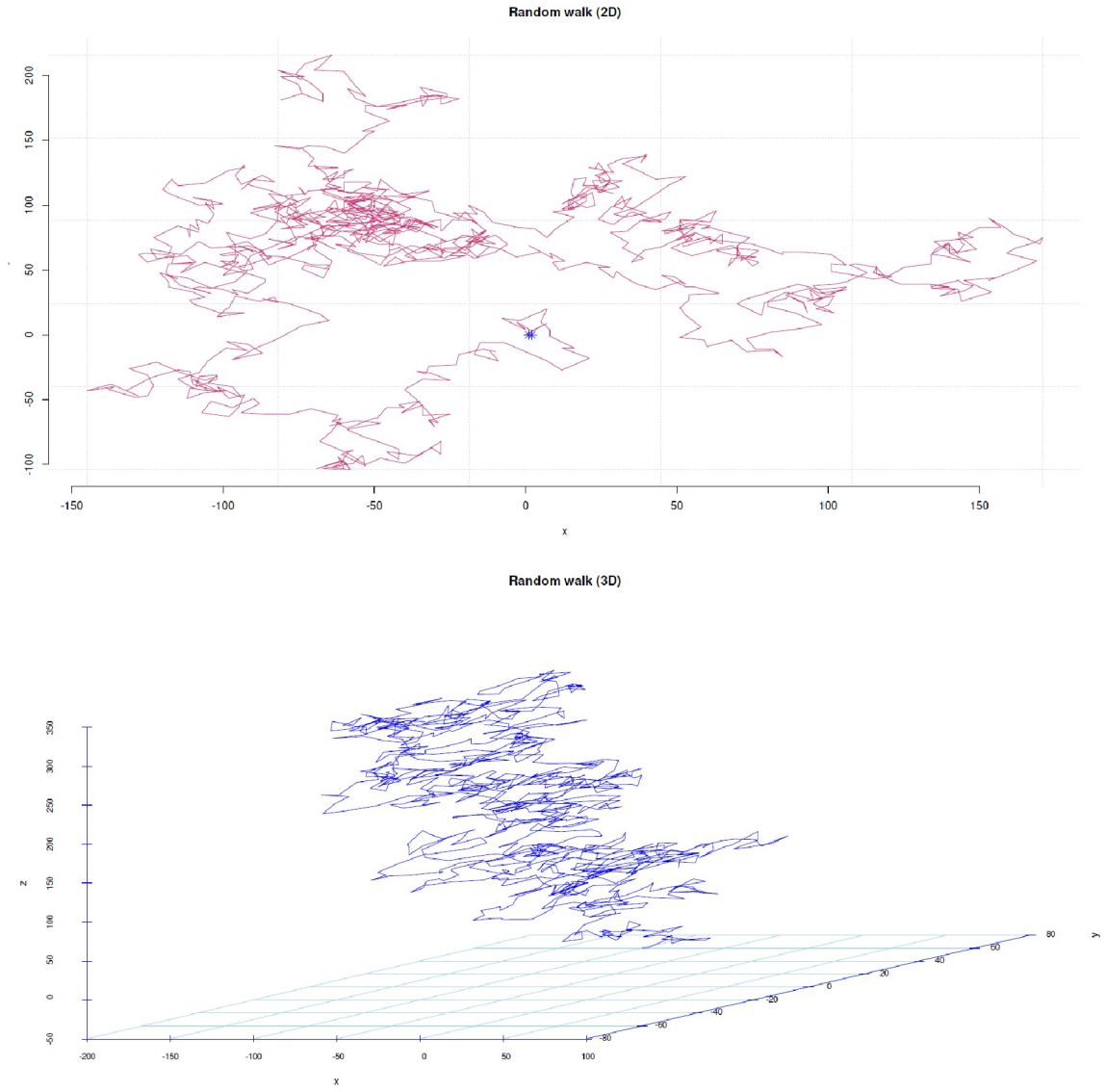


Figura 6.78. *Cadena de Markov (Random Walk en 2D y 3D)*

La simulación de π es tan breve y elegante que merece la pena imprimir primero el código R de Eddelbüttel (ibid.). Luego examinamos lo que el algoritmo realmente hace (`ptII_Bayes_quan_simulate--pi.r`):

```
> seed <- 5
> set.seed(seed)
> for(i in 10^(1:7))
+ {
+ cat("i = ",i,"\t",piR(i),"\n")
+ }
i = 10 3.2
i = 100 2.96
i = 1000 3.092
i = 10000 3.1544
i = 1e+05 3.13892
i = 1e+06 3.141028
i = 1e+07 3.141044
```

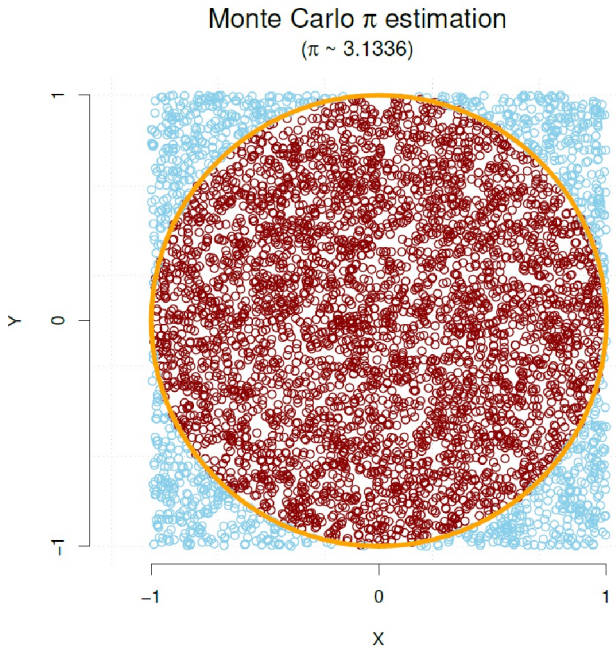


Figura 6.79. *Cálculo de* π

La simulación se basa en el cálculo del área de un círculo que se dispone dentro de un cuadrado de forma que los lados del cuadrado estén tangencialmente tocados por el círculo (véase la Fig. 6.79). El área conocida del círculo A_K se calcula a partir de π el radio r para ser $A_k = \pi * r^2$. El diámetro es d = 2* r, de lo que se deduce para el área

$$A_k = \pi \cdot \left(\frac{d}{2}\right)^2 \tag{6.109}$$

$$= \frac{1}{4} \cdot \pi \cdot d^2 \tag{6.110}$$

Esto se puede resolver π con $\pi = 4*A_k/d^2$: El área del cuadrado se calcula a partir de A_Q = x*y, en el caso del cuadrado como $A_Q = x^2$, expresado con el radio del círculo $A_Q = d^2 = (2*r)^2$. El cociente de las áreas de A_K y A_Q es

$$R_{A_K \longleftrightarrow A_Q} = \frac{A_K}{A_Q} \tag{6.111}$$

$$= \frac{\pi \cdot r^2}{2 \cdot r^2} \tag{6.112}$$

$$= \frac{\pi}{4} \tag{6.113}$$

Ahora se generan aleatoriamente N puntos dentro del cuadrado mediante simulación y se comprueba si forman parte del área circular o se encuentran fuera. Normalmente, hay S = N*π /4 puntos dentro del círculo, es decir, dentro su área. Por tanto, la cantidad buscada puede aproximarse a

$$\pi = \frac{4 \cdot S}{N} \tag{6.114}$$

y se busca el número de puntos dentro del círculo. Un punto de coordenadas (X,Y) se encuentra dentro del círculo si se cumple lo siguiente

$$(x^2 + y^2) < r^2 \tag{6.115}$$

Se cuenta el número de puntos y se introduce el resultado en la fórmula para π. Cuantos más puntos, más exacto será el resultado. Hay que reconocer que el procedimiento es bastante ineficaz y lento porque utiliza toda el área del cuadrado, es decir, el área del círculo y el área exterior a él. Sin embargo, funciona y es fácil de aplicar, pero supera la capacidad del ordenador para un mayor número de decimales debido a la programación. Otros métodos no utilizan el área del círculo sino la circunferencia para aproximar. Esto es más eficaz. Entonces hay sin duda un buen número de algoritmos especiales para llegar a los buenos 50 billones de decimales en un futuro previsible, que es el actual récord mundial establecido por Timothy Mullican (2021) en 2020. Se trata aquí de demostrar que, además de la sencillez de generar simulaciones con los algoritmos adecuados, se puede sacar provecho de extraer números aleatorios de una distribución uniforme. Posteriormente, en el algoritmo Metrópolis-Hastings, se utiliza un valor aleatorio de la distribución uniforme para medir si un valor generado por el algoritmo representa (o no) una mejora en el contexto del algoritmo MCMC. De este modo, el espacio de parámetros se recorre sucesivamente de forma simulativa.

El bootstrapping (Efron, 1979) y las pruebas de permutación (Fisher, 1935/1973) también forman parte de las simulaciones MC, al igual que la generación de distribuciones predictivas posteriores (Gelman, Carlin, Stern & Rubin, 2004, cap. 11). La tabla de Galton también representa una simulación MC en la que la distribución binomial se forma después de muchas repeticiones. El siguiente código R (véase la Fig. 6.80, arriba) simula tanto ésta como una distribución normal y traza la distribución de densidad exacta teórica (`ptII_quan_Bayes_MC-simulation_binom-norm.r`) para comparalas. La figura (6.80, abajo) contiene lo mismo para la distribución normal (código R no impreso).

```
# MC-Simulation
# binomial distribution
set.seed(223)
N <- 1e+6
p <- 0.5
size <- 20
rb <- rbinom(n=N, size=size, p=p)
db1 <- density(rb)
sek <- 1:size
db2 <- dbinom(x=sek,size=size,p=p)
fac <- 1.12
ylim <- range(c(db1$y, db2))*c(1,fac)
xlim <- c(0,size)
rb.dens <- density(rb)
ylim <- c(0,max(rb.dens$y))
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l")
```

```
hist(rb, ylim=ylim, xlim=xlim, prob=TRUE, border="white",
       col="skyblue",ylab="Density", pre.plot=grid(), xlab="size", main="")
lines(sek,db2, col="darkred", lty=2, lwd=2, type="h")
lines(sek,db2, col="darkred", lty=2, lwd=2, type="l")
lines(rb.dens, col="blue", lty=2, lwd=2, type="h")
mtext("Monte Carlo Simulation of a binomial density",
       outer=TRUE, line=-1.4, cex=1.5, side=3)
mtext(eval(substitute(expression(paste("N = ", N, " | p = ",p," | size = ",size)),
       list(N=N, p=p, size=size))), outer=TRUE, line=-3.3, cex=1, side=3)
```

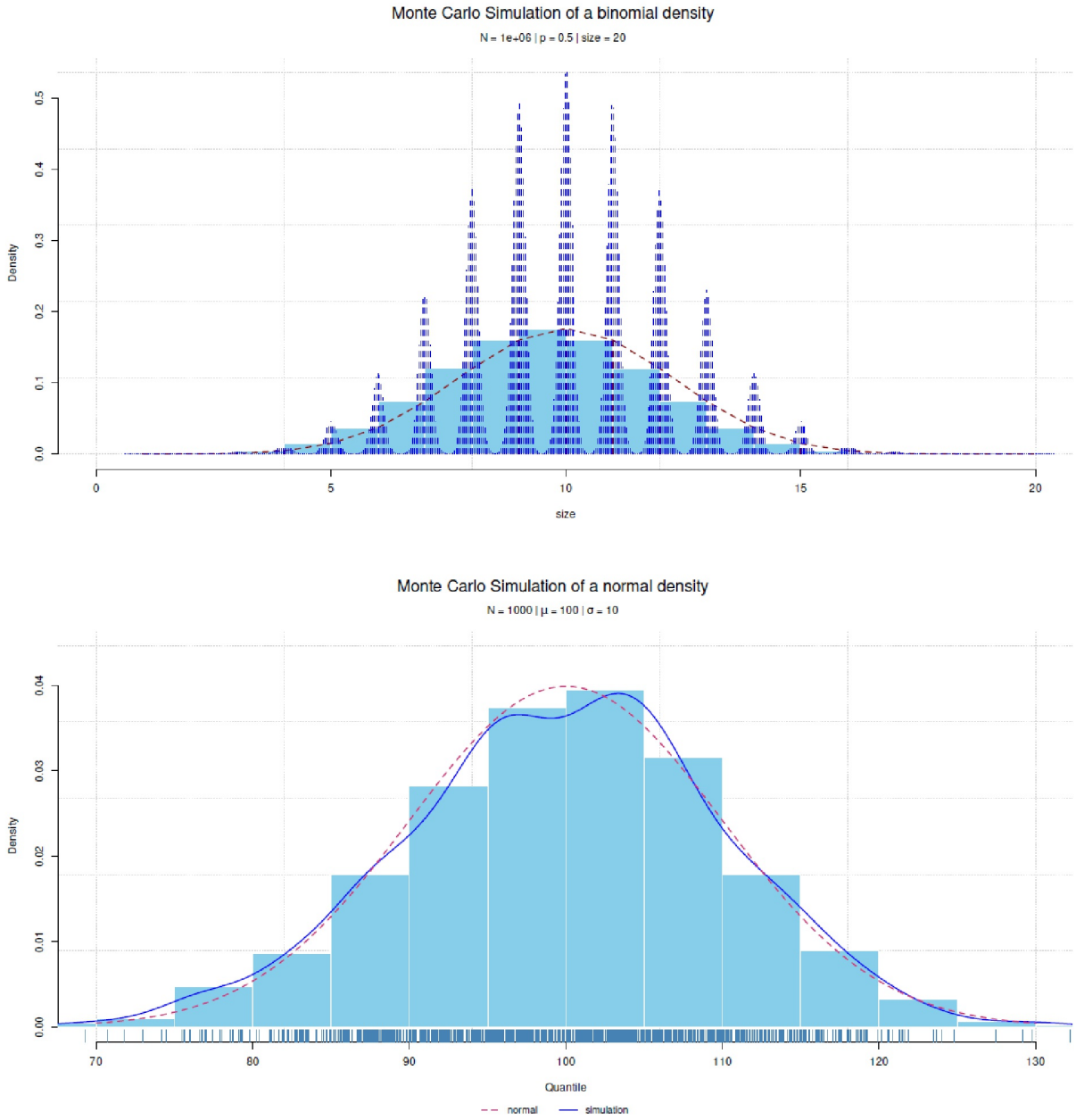


Figura 6.80. *Densidad de simulación Monte Carlo (modelo de distribución binomial y normal)*

Si ahora combinamos las cadenas de Markov con simulaciones MC, surge algo apasionante: podemos simular un modelo, por ejemplo, una ecuación de regresión. Podemos simular un modelo, por ejemplo una ecuación de regresión, y asignar una distribución de probabilidad a cada parámetro, estimando así todo el

modelo. A partir de distribuciones a priori por parámetro, se toman muestras a lo largo de la ecuación del modelo (aquí sería la regresión) extraídas de tal manera que el proceso se desarrolle de extracción en extracción según los principios de las cadenas de Markov. Las extracciones representan muestras aleatorias del conjunto del modelo considerado. El proceso de simulación puede optimizarse en función de determinados criterios, como *caminar* únicamente a lo largo de gradientes ascendentes (Hamilton-Monte Carlo, véase más adelante) o aceptar únicamente valores según determinados criterios, como en el algoritmo Metropolis-Hastings (véase más adelante). Se puede complicar estos criterios tanto como quiera. Pero básicamente se trata de la regla de asignación lo más eficiente posible, de cómo recorrer aleatoriamente el espacio de parámetros de un modelo multidimensional completo con el fin de cartografiar todo el espacio lo más rápidamente posible y obtener una estimación robusta de todos los parámetros relevantes a lo largo de la ecuación del modelo dado.

Los algoritmos MCMC difieren en cuanto a su eficacia y manejabilidad para converger de forma estacionaria, es decir, para llegar a una solución plausible o en absoluto, incluso en el caso de modelos complejos, y de tal forma que las muestras aleatorias que entran en el pool sean independientes entre sí. Se puede investigar esto (véase también el capítulo 6.13.3 sobre el diagnóstico de las cadenas MCMC) mediante un gráfico de autocorrelación `acfplot()`, un traceplot y el histograma, véase los paquetes R `coda` y `bayesplot`. Aunque hay muchos subtipos de algoritmos MCMC (Clark, 2016), se pueden distinguir tres algoritmos que son particularmente comunes: *Metropolis-Hastings, Gibbs Sampling* y *Hamilton Monte Carlo*. Se pueden encontrar más detalles sobre los algoritmos en Kruschke (2015b), Gelman, Carlin, Stern y Rubin (2004), Bolstadt (2007) o McElreath (2015). (2015), por nombrar solo algunos libros comunes.

6.13.1 Los algoritmos MCMC

6.13.1.1 Algoritmo Metropolis-Hastings

El algoritmo de Metrópolis-Hastings (= MH) se basa en el algoritmo de Metrópolis (Metrópolis, Rosenbluth, Rosenbluth, Teller & Teller, 1953; Metrópolis & Ulam, 1949; Metrópolis, 1987). Según Wikipedia (2019j), es posible que el algoritmo se remonte a la pareja Rosenbluth y que Metrópolis se llevara erróneamente los créditos por él. Quizá la parte de Ulam también sea mayor de lo que se piensa, como señala Gelman (2014e). El algoritmo de Metrópolis genera una cadena de Markov y, por tanto, los estados del sistema según la distribución de Boltzmann. Se le utilizó originalmente para simular la distribución de los estados del sistema de moléculas idealizadas. Al principio, el movimiento molecular se ajustaba a la mecánica clásica, en honor al matemático inglés, físico y astrónomo Isaac Newton (1642-1726/27) y fue reformulado matemáticamente por la mecánica hamiltoniana en honor al matemático irlandés William Rowan Hamilton (1805-1865). Esta teoría tomó como punto de partida la mecánica clásica y a la larga desembocó en la mecánica estadística y la mecánica cuántica. En concreto, Newton trabajaba con vectores, mientras que la reformulación de las leyes del movimiento utiliza funciones energéticas de objetos en interacción. Hastings (1970) amplió el algoritmo de Metrópolis y lo generalizó de forma que ahora las variables aleatorias se simulan mediante cadenas de Markov, de forma que se puede simular cualquier distribución deseada. Esto es especialmente útil si la distribución deseada no se puede simular directamente y no se pueden resolver las integrales analíticamente de todos modos. El algoritmo supone que la densidad de la distribución deseada puede calcularse en todo el espectro de la distribución. Trasladado a los análisis estadísticos, esto significa que se pueden simular modelos bayesianos complejos, como las distribuciones posteriores de los parámetros del modelo de interés. Éstas se pueden analizar posteriormente o se pueden derivar de la simulación variables cuantitativas de interés. El algoritmo MH genera muestras de la distribución de probabilidad compleja para cada variable de interés. Los candidatos se generan aleatoriamente a partir de la distribución de probabilidad. Sin embargo, el algoritmo no acepta automáticamente estos candidatos. La distribución de probabilidad objetivo puede ser simétrica o asimétrica. Las distribuciones asimétricas resultan cuando hay limitaciones (por ejemplo, no se permiten valores menores que cero para las varianzas) o la propia distribución es sesgada (por ejemplo,

la distribución lognormal). La distribución de probabilidad usualmente no es definitiva y no se puede determinarla más allá de toda duda. En la gran mayoría de los casos, quedan grados de libertad o una cierta incertidumbre sobre la forma real. Una regla general es que la aceptación de candidatos no debe ser extrema, es decir, ni *nunca* ni *siempre y todo*. Hay que desconfiar de las tasas de aceptación extremas, que indicarían que la distribución de la densidad objetivo es demasiado estrecha o demasiado amplia.

Se pueden generar los puntos de partida, es decir, los valores de entrada para las simulaciones, a partir de las distribuciones a priori de los parámetros o especificarlos manualmente en cada caso. La probabilidad de aceptación resulta del cociente de la probabilidad posterior del candidato *i* frente a la del candidato *i -1* de la simulación anterior. El candidato *i* es aceptado si la probabilidad de aceptación es mayor que la de la variable aleatoria de la distribución uniforme. Entonces el candidato *i* sustituye al candidato anterior *i-1*. Esto garantiza que el muestreador (sampler) cubra las zonas de la distribución de probabilidad objetivo con probabilidades más altas. Al no aceptar todos los rangos, el algoritmo tiene un límite natural incorporado para deshacer una dirección errónea (por ejemplo, donde los valores podrían atascarse, por así decirlo) y realinearse (Gilks, Richardson & Spiegelhalter, 1996). Tras decidir si se acepta o no al candidato, se repite el procedimiento con los valores objetivos así actualizados – y tiene lugar la siguiente simulación. Aunque el procedimiento no es el más eficiente, ya que se ejecuta con bastante lentitud como un paseo aleatorio a través del espacio de parámetros, se le considera muy robusto y generalmente conduce al objetivo. El objetivo es replantear de toda la distribución de probabilidad de interés. La tasa de aceptación y los distintos gráficos de las simulaciones MCMC (véanse 6.13.2.1 y 6.13.2.3.3 con ejemplos de diagnósticos MCMC) dan una impresión del curso del muestreo. El procedimiento genérico del algoritmo MH es el siguiente

1. Definición de la función objetivo *F*
   - Establecer una función objetivo *F*, lo que se va a simular (por ejemplo, media, varianza, modelo de regresión, etc.).
     - El algoritmo MH produce una distribución de muestreo condicional a esta función objetivo *F*.
     - En modelos más complejos, se incluyen Prior y Likelihood aquí, de modo que la función objetivo *F* genera la distribución posterior objetivo, es decir, *F = Posterior = Prior * Likelihood*. Entonces se necesitan funciones separadas para laPrior y la Likelihood.

2. Elección de un valor inicial
   - Esto puede corresponder a una estimación cualitativa. La elección debe ser razonable y no demasiado extrema, pero dentro del rango de valores posibles. Sin embargo, es sólo el punto de partida.
   - A menudo, las primeras *z* iteraciones se omiten para las estimaciones posteriores y se consideran estas iteraciones como secuencias de "quemar".

3. Cadena de Markov: Inicio de una cadena de Markov con i_j iteraciones
   - Paso de propuesta
     - Almacenar el valor actual x_j*-1* de la cadena de Markov.
     - Calcular la propuesta para el valor futuro, el candidato $x^{cand}_j = x_j\text{-}1$ + parte aleatoria. Esto corresponde al muestreo x^{cand}_j de la distribución distribución propuesta.
   - Paso de evaluación o aceptación
     - Insertar x^{cand}_j y x_j-1 en la función objetivo F.
     - Calcular la relación *R* con $R = x^{cand}_j/x_j\text{-}1$
     - En caso de que $R <$ unif(0,1), es decir, *R* sea menor que un número aleatorio de la distribución uniforme en el intervalo de cero a uno, entonces se toma x^{cand}_j como nuevo valor para la siguiente iteración j con probabilidad min(1,R). Sustituye al anterior, es decir, $x_j = x^{cand}_j$. En caso contrario, el candidato es rechazado y se mantiene el valor actual, es decir, $x_j = x_j\text{-}1$.
   - Guardar x_j también puede hacerse sólo cada *m* iteraciones para mantener las dependencias de los sorteos (=paseo aleatorio/random walk por el espacio de parámetros) entre sí (autocorrelaciones).
   - Volver al punto "paso de propuesto, siguiente iteración, y continuar con la cadena de Markov hasta que se hayan ejecutado todas las simulaciones *i* previstas.

4. Diagnóstico
   - Al final, se pueden evaluar los valores simulados *x* (distribución muestral de *F*) mediante diagnósticos MCMC apropiados y se pueden sacar correspondientes conclusiones relacionadas con *F* mediante estimaciones de parámetros, convergencia de las simulaciones MCMC, etc.

6.13.1.2 Muestreo de Gibbs

El muestreo de Gibbs (Casella & George, 1992) es un caso especial del algoritmo Metrópolis-Hastings en el que la distribución de densidad objetivo corresponde a la distribución MCMC posterior resultante. Técnicamente, cada candidato se acepta como representante legítimo de la distribución de densidad objetivo. Por tanto, la probabilidad de aceptación es siempre p = 1. Si la muestra se elige lo suficientemente grande, cualquier característica de la población o la propia distribución de densidad puede simularse con cualquier grado de precisión (ibíd., p.168). Aunque los cálculos se basan en simulaciones los resultados finales representan las cantidades de la población. Básicamente se toma una muestra de una función *f(x)*, que representa una densidad marginal de una distribución conjunta *(marginal density de una joint distribution*), sin que *f(x)* sea directamente necesaria. Este es el caso, por ejemplo, cuando las integrales subyacentes no pueden calcularse fácilmente de forma analítica o numérica. Dadas son así (ibíd., p.167s.) la densidad conjunta (*joint density*)

$$f(x, y_1, \ldots, y_p) \tag{6.116}$$

y la densidad marginal (*marginal density*)

$$f(x) = \int \ldots \int f(x, y_1, \ldots, y_p)\, \mathrm{d}y_1 \ldots \mathrm{d}y_p \tag{6.117}$$

y la media o la varianza. El muestreador de Gibbs sustituye así al cálculo directo de *f(x)*. En el caso simple de dos variables con un par de variables aleatorias (X,Y), se genera una *secuencia de Gibbs* de variables aleatorias

$$Y_0', X_0', Y_1', X_1', Y_2', X_2', \ldots, Y_k', X_k' \tag{6.118}$$

Se determina el valor inicial $Y'_0 = y'_0$ y todos los demás valores de la secuencia de Gibbs son generados por el algoritmo de simulación alternando entre X e Y según

$$X_j' \sim f(x \mid Y_j' = y_j') \tag{6.119}$$

$$Y_{j+1}' \sim f(y \mid X_j' = x_j') \tag{6.120}$$

En caso de que sea $k \to \infty$ convierte X'_k a *f(x)*, la "verdadera" distribución marginal de *X*. Si *k* es suficientemente grande, el dato final es $X'_k = x'_k$, donde *k* es un punto muestral efectivo de *f(x)*. A partir de él, se puede *reconstruir* qua simulación toda la distribución muestral buscada de *f(x)*. También, si *k* es suficientemente grande, se pueden *generar* extracciones independientes de la distribución muestral de *f(x)*. Lo que funciona en el caso bivariante simple es igualmente posible en el caso multivariante, por tanto más complejo, ya que la simulación se realiza a partir de integrales multidimensionales (ibíd., p. 172). La distribución de muestreo generada de este modo puede examinarse diagnósticamente de la forma habitual (véase el capítulo 6.13.2.2), por ejemplo, para comprobar la convergencia y la independencia de las extracciones y calcular posteriormente las cantidades de interés. El seguimiento de la secuencia de Gibbs es relevante para encontrar un valor adecuado para *k*. El muestreador de Gibbs resulta más útil cuanto más compleja es la distribución que se va a simular y cuantas más dimensiones haya que incluir, que el muestreador de Gibbs puede descomponer en determinaciones numéricas en dimensiones inferiores. Esto

es práctico y pragmático. La eficacia de la muestra de Gibbs, por otra parte, depende de la tasa de convergencia, de modo que la distribución empírica de muestreo $X´_j$ converge a *f(x)*. Cuanto más rápido se mueve $X´_j$ por el espacio muestral, más rápida será la velocidad de convergencia. El muestreador de Gibbs puede utilizarse tanto para el cálculo de las distribuciones a posteriori bayesianas como para el cálculo clásico de las funciones de Likelihood y las propiedades de los estimadores de Likelihood.

6.13.1.3 Hamilton Monte Carlo

El Hamilton Monte Carlo (= HMC), originalmente llamado Hybrid Monte Carlo (Alder & Wainwright, 1959; Neal, 2011; Betancourt, 2017; Duane, Kennedy, Pendleton & Duncan, 1987; Homan & Gelman, 2014), es en definitiva otro caso especial – o más bien un desarrollo del algoritmo MH, que funciona de forma mucho más eficiente que este último. El punto de partida es una interpretación física del algoritmo con referencia directa a la mecánica estadística. La exploración del espacio de parámetros se modela como un movimiento sin fricción a lo largo del gradiente de la función objetivo. El movimiento viene determinado por la posición y el momento. Este enfoque físico se combina con la evaluación de los candidatos según el algoritmo MH.

En su interpretación física, se puede entender el algoritmo HMC de tal manera que se da una bola de billar y un taco a un algoritmo MH y en lugar de recorrer lenta y laboriosamente el espacio de parámetros, en función del gradiente y otros factores de la bola de billar, se realiza un pequeño empujón con el taco de tal manera que la bola de billar rebote un poco sin pérdida de fricción. En estadística este proceso se equipara a menudo con el salto de rana (= leapfrog) y el paso del proceso se denomina "paso de salto de rana" (= *leapfrog step*). El leapfrog es el elemento especial del algoritmo HMC. Partiendo de valores aleatorios estándar normalmente distribuidos para los parámetros a estimar, este leapfrogging se convierte ahora en un candidato (vector) para la cadena MCMC de acuerdo con ciertas reglas (Neal, 2011; Betancourt, 2017), de modo que se llega a un punto final del movimiento de leapfrogging. Al hacerlo, el gradiente de la función objetivo entra en los cálculos. Este nuevo punto es el siguiente (vector de) candidato(s) del algoritmo HMC. Lo que viene a continuación corresponde a la evaluación habitual del algoritmo HMC del candidato (vector). Éste se compara con un número aleatorio distribuido uniformemente, es decir, si la probabilidad de este nuevo candidato (vector) es superior al número aleatorio generado. En caso afirmativo, el candidato (vector) se incluye como siguiente valor de comparación (o vector) y constituye el punto de partida de la siguiente repetición. Todo el proceso se repite hasta que se ha saltado a través del espacio de parámetros, es decir, se ha realizado el número de repeticiones exitosas necesarias para obtener estimaciones suficientemente precisas de los parámetros apuntados.

Los fundamentos y el concepto de HMC están excelentemente descritos por Betancourt (2017) y Neal (2011), respectivamente. Seguimos la terminología de Neal (ibíd.). La dinámica hamiltoniana subyacente describe la evolución de los sistemas físicos dinámicos como un formalismo matemático. En física, el enfoque se utiliza para describir, por ejemplo, sistemas planetarios o de electrones en un campo electromagnético. Con referencia a la simulación de distribuciones de probabilidad dirigidas, Neal (ibíd.) establece la comparación con un disco en el hockey sobre hielo sin fricción, que se desliza sobre la superficie de la distribución en dos dimensiones sin pérdidas por fricción y puede moverse a diferentes alturas. Son relevantes la posición del disco, descrita por un vector bidimensional *q* y el momento del disco, que corresponde a su masa multiplicada por la velocidad y está descrito por el vector bidimensional *p*. La energía potencial *U(q)* del disco es proporcional a su altura sobre la superficie en su posición actual. La energía cinética del movimiento *K(p)* es igual al momento *p* al cuadrado dividido por la masa *m* del disco. En el caso más sencillo, se supone una masa $m = 1$, de modo que *m* desaparece de la ecuación.

$$K(p) = \frac{p^2}{2 \cdot m} \qquad (6.121)$$

La velocidad del disco en la superficie corresponde a un movimiento constante *p*/*m* con la fuerza -m*g, donde *g* es la aceleración debida a la gravedad. Si el gradiente aumenta, el impulso permite que la velocidad continúe con energía cinética decreciente y energía potencial creciente. Si $p = 0$, el movimiento se detiene

y el disco se desliza hacia atrás con energía cinética creciente y energía potencial decreciente. Aparte de la suavidad del movimiento, este comportamiento corresponde exactamente a lo observado en la realidad física.

Aplicado a la estadística, la posición corresponde a las variables de interés. La energía potencial *U* viene dada como la log-probabilidad negativa de la distribución de probabilidad de interés (= modelo estadístico investigado). Si π*(q)* es la densidad a priori y *L(q|D)* es la función de Likelihood dados los datos, se dice que la energía potencial *U* es es un producto de la función de la Prior y la función de la Likelihood:

$$U(q) = -\log[\pi(q) \cdot L(q \mid D)] \tag{6.122}$$

Hay variables de impulso artificiales por cada variable de posición. El número de dimensiones examinadas de la dinámica hamiltoniana se compone del vector d-dimensional de posición *q* y del vector d-dimensional de momento *p*, de modo que hay un total de 2*d dimensiones. La función de Hamilton *H(q,p)* describe entonces completamente todo el sistema. La energía total *E* es entonces consecuencia de la combinación de la energía potencial *U(q)* y la energía cinética *K(p)*.

$$H(q, p) = U(q) + K(p) \tag{6.123}$$

y la energía total

$$E = E(q, p) \tag{6.124}$$

A partir de aquí, se pueden derivar ecuaciones parciales de movimiento, que determinan el comportamiento de *q* y *p* a lo largo del tiempo *t* con respecto a las ecuaciones de Hamilton. Para las ecuaciones de Hamilton con índice i = 1, ..., d entonces se aplica

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i} \tag{6.125}$$

$$\frac{dp_i}{dt} = \frac{\partial H}{\partial q_i} \tag{6.126}$$

A partir de ahí, se pueden determinar los estados por tiempo *t* con duración *s* para *t+s*, de modo que ambos juntos produzcan un mapa o función de asignación T_s. El tiempo *t* se elige para integrar el paso de salto leapfrog numéricamente en el tiempo. Para el álgebra matemática, *q* y *p* pueden combinarse en un vector *z = (q, p)* con dimensiones *2*d* y, por tanto, juntar las ecuaciones. La ecuación de Hamilton se transforma entonces en

$$\frac{dz}{dt} = J \cdot \nabla H(z) \tag{6.127}$$

donde ∇*H* es el gradiente de *H*

$$[\nabla H]_k = \frac{\partial H}{\partial z_k} \tag{6.128}$$

Para *J* (2*d x 2*d matriz, definición de cuadrantes por identidad y cero-matrices) se aplica

$$J = \begin{bmatrix} 0_{d \times d} & I_{d \times d} \\ I_{d \times d} & 0_{d \times d} \end{bmatrix} \tag{6.129}$$

Las reglas de la dinámica hamiltoniana especifican así cómo procede exactamente el movimiento de rebote a través del espacio de parámetros de la distribución de probabilidad que se desea explorar. Los cálculos funcionan con derivadas parciales en función de *p* y *q* a lo largo del tiempo *t*. Si no se piensa en estadística, sino en física y mecánica, el procedimiento no es tan difícil de entender. Especialmente en el espacio de alta dimensión, las matemáticas se vuelven rápidamente complejas y pueden fallar. De todos modos, las soluciones analíticas están fuera de lugar para casi todos los problemas complejos. El algoritmo HMC ofrece una posibilidad práctica de alcanzar un objetivo en un tiempo previsible de forma metódicamente controlada.

6.13.1.3.1 Características de la dinámica hamiltoniana – El enfoque físico elegido de la dinámica hamiltoniana muestra varias características que se consideran ventajas (Neal, 2011, p.116.):

- *Reversibilidad* – La dinámica hamiltoniana es reversible en el sentido de que los pasos pueden en principio retroceder en el tiempo *t* sin provocar cambios alejados de la expectativa de retroceso en el tiempo. El mapa T_s corresponde a *(q(t), p(t)) → (q(t+s), p(t+s))* con la inversa T_{-s}, es decir, negación de *p*, aplicación de T_s y negación de *p* de nuevo. Se supone que *H* y T_s no dependen de *t*, es decir, que son invariantes en el tiempo. La distribución de probabilidad apuntada y, por tanto, las actualizaciones MCMC no provocan ningún cambio. Todas las transiciones de la cadena de Markov son reversibles.
- *Preservación de la función hamiltoniana* – El hamiltoniano *H(q, p)* se mantiene invariante, es decir, se conserva o preserva. La puntuación de aceptación de Metropolis es 1 si *H* se mantiene perfectamente invariante. En la práctica, sin embargo, este valor no se alcanza y *H* es aproximadamente invariante. Así pues, *H(p, q)* es invariante a lo largo del tiempo *t*.
- *Invarianza del volumen* – El mapa T_s mantiene constante el volumen en el espacio *(q, p)*. Incluso si una región se estira en una dirección por la aplicación de la dinámica hamiltoniana, simultáneamente se comprime en otra dirección en la misma cantidad, de modo que en conjunto el volumen permanece constante. Esto tiene implicaciones para el procedimiento de aceptación y la probabilidad de aceptación en la actualización de Metrópolis, que no requiere ningún ajuste.
- *Estructura simpléctica* – La geometría simpléctica permite formular movimientos en mecánica mediante ecuaciones sin coordenadas y se utiliza especialmente en la mecánica hamiltoniana. Con ella, la dinámica puede explorarse más fácilmente. La condición simpléctica es válida si *z = (q, p)* y *J* se definen como arriba, de modo que la matriz de Jacobi *B* de las derivadas de las T_s asociadas satisface la siguiente ecuación:

$$B_s^T \cdot J^{-1} \cdot B_s = J^{-1} \tag{6.130}$$

Esto implica la conservación del volumen (véase más arriba), ya que la ecuación también se aplica a los determinantes de los elementos de la ecuación:

$$J^{-1} \cdot \det(B_s^T) \cdot \det(J^{-1}) \cdot \det(B_s) = \det(J^{-1}) \tag{6.131}$$

y eso implica

$$\det(B_s)^2 = 1 \tag{6.132}$$

Es importante señalar que estas características siguen siendo válidas incluso cuando se aproxima la dinámica hamiltoniana, que corresponde a la aplicación en la realidad y permite ciertas tolerancias o desviaciones.

6.13.1.3.2 El paso de leapfrog – Para calcular los impactos cinéticos, se aplican las leyes de la cinética: la energía potencial *U()* en la posición actual *p* da *U(q)* y la energía cinética *K()* con su momento *p* da *K(p)*. Juntas forman el sistema de Hamilton. La solución de las ecuaciones parciales puede realizarse por el método de Euler (= integración numérica) con índice $i = 1,...,d$ y comenzar en $t = 0$ con los valores iniciales $q_i(0)$ o $p_i(0)$

$$p_i(t+\epsilon) = p_i(t) + \epsilon \cdot \frac{dp_i}{dt}(t) \tag{6.133}$$

$$= p_i(t) - \epsilon \cdot \frac{\partial U}{\partial q_i}(q(t)) \tag{6.134}$$

$$q_i(t+\epsilon) = q_i(t) + \epsilon \cdot \frac{dq_i}{dt}(t) \tag{6.135}$$

$$= q_i(t) - \epsilon \cdot \frac{p_i(t)}{m} \tag{6.136}$$

o con el algoritmo leapfrog. Una mejora del planteamiento anterior es hacer tres pasos en lugar de dos y dividir el primer paso de Euler *antes* y *después* de la actualización del impulso:

$$p_i(t+\frac{\epsilon}{2}) = p_i(t) + \frac{\epsilon}{2} \cdot \frac{\partial U}{\partial q_i}(q(t)) \tag{6.137}$$

$$q_i(t+\epsilon) = q_i(t) + \epsilon \cdot \frac{p_i(t+\frac{\epsilon}{2})}{m_i} \tag{6.138}$$

$$q_i(t+\epsilon) = p_i(t+\frac{\epsilon}{2}) - \frac{\epsilon}{2} \cdot \frac{\partial U}{\partial q_i}(q(t+\epsilon)) \tag{6.139}$$

El leapfrog tiene entonces las secuencias de paso:

1. Comienza con medio paso para el momentum,
2. luego un paso entero para la posición usando los nuevos valores para el momento y
3. termina con otro medio paso para el momento utilizando los nuevos valores para la posición.

Los pasos 1 a 3 pueden repetirse para ir de $t+\epsilon$ a $t+2\epsilon$. A continuación, se puede combinar el primer medio paso de la primera actualización de $p_i(t + \epsilon/2)$ a $p_i(t + \epsilon)$ con la primera mitad de la segunda actualización de $p_i(t + \epsilon)$ a $p_i(t + \epsilon + \epsilon/2)$. Esto da la regla para múltiples pasos de leapfrog, que corresponde a la práctica. Esto se debe a que en la práctica se realizan varios pasos de leapfrog antes de hacer una propuesta. A esto le sigue el último paso de aceptación de Metropolis para probar la propuesta (= candidato/vector). Dado que se trata de puras transformaciones de datos, todos los pasos pueden invertirse y, en principio, son reversibles. El procedimiento es preciso, ya que según Neal (2011, p.122) el error para (q, p) o $H(q, p)$ no es superior a ϵ^2 / y tiende a cero con $\epsilon \to 0$.

En general, el paso de leapfrog conduce a una reducción significativa de las correlaciones entre los estados extraídos con éxito de la distribución de probabilidad esperada. Esto significa que las extracciones son más independientes entre sí en comparación con el algoritmo MH, ya que hay un mayor espacio o distancia entre las extracciones. Esto significa que la distancia entre las muestras generadas suele ser muy grande. Además, la atención se centra en tasas de aceptación elevadas. Esto aumenta el esfuerzo por simulación, ya que el salto se divide en diferentes subpasos de integración numérica. Pero la secuencia converge más rápidamente, de modo que, en conjunto, el procedimiento alcanza su objetivo con mayor rapidez y, aun así, de forma segura. El algoritmo MH, por otra parte, explora el espacio de parámetros con bastante lentitud y no muy eficientemente (= mayores tasas de no aceptación) a través de los simples candidatos a paseo aleatorio de la distribución de probabilidad indicada.

6.13.1.3.3 La evaluación de aceptación de Metropolis – Los candidatos propuestos, los nuevos vectores q y p, se comparan con un número aleatorio distribuido uniformemente y se les asigna la probabilidad

$$\mathrm{MH}_{\text{crit}} = \min[1, \exp(-H(q^*, p^*) + (Hq, p))] \quad (6.140)$$

$$= \min[1, \exp(-U(q^*) + U(q) - K(p^*) + K(p))] \quad (6.141)$$

aceptada. Esto determina si los candidatos se utilizan como nuevo patrón de comparación en la siguiente ejecución o si se conservan los anteriores. Si se aceptan, los nuevos candidatos se añaden a la distribución de probabilidad apuntada y el bucle comienza desde el principio como antes hasta que se completan las repeticiones MCMC apuntadas y se ejecuta en paralelo durante la duración de las cadenas MCMC apuntadas cada una por separado. Neal (2011) resume todo el proceso de HMC en unas pocas líneas de código R (véase la Sección 6.13.2.3.1).

6.13.1.3.4 La elección de los parámetros de salida y los problemas – La elección del tamaño del paso y el número de saltos leapfrog *L* resultan no ser triviales en la práctica. Así, los pasos pueden elegirse demasiado grandes o demasiado pequeños, con lo que no se avanza en el espacio de parámetros o se acaba fuera de él o en los extremos y se pierde allí. O bien el espacio de parámetros se recorre muy lentamente hasta que se alcanzan valores estables o bien hay una alta tasa de rechazo y sólo bajas tasas de aceptación de los candidatos. Lo mismo ocurre con el número de saltos. Por un lado, esta optimización no debe llevar demasiado tiempo, pero tampoco demasiado poco, para llegar a candidatos utilizables. Además, estos parámetros cambian en función del modelo y de los datos, y no se puede afirmar que una determinada combinación de tamaño de paso y número de leapfrogs sea siempre óptima en todas las circunstancias.

Otro problema puede surgir si el espacio de parámetros apuntados tiene largas colas en los extremos, de modo que el algoritmo pasa mucho tiempo en los valores extremos y las zonas más esenciales del espacio de parámetros quedan fuera. Betancourt (2017) discute estos y otros problemas y señala diagnósticos para encontrar tales problemas numéricamente o gráficamente. El autor hace hincapié en que la investigación teórica de la distribución apuntada siempre es necesaria para detectar desviaciones patológicas. Esto es aún más cierto para todos los problemas complejos. Sin embargo, esto no disminuye la bondad del HMC, ya que otros enfoques muestran problemas comparables. Así pues, el algoritmo HMC se considera un procedimiento robusto cuyo resultado puede examinarse bien desde el punto de vista del diagnóstico. La posibilidad de ejecutar cadenas MCMC paralelas y compararlas entre sí aumenta la transparencia y, por tanto, la confianza en el procedimiento.

Ninguno de los problemas mencionados se plantearía en la práctica. En la práctica con R (s. cap. 6.13.2.3) mostramos mediante un pequeño ejemplo cómo problemas como el tamaño del paso pueden realmente tener un efecto. Sin embargo, dado que la determinación de los valores óptimos para el tamaño del paso y el número de saltos consume mucho tiempo en la práctica, los autores están intentando encontrar soluciones automáticas en el desarrollo posterior del HMC.

6.13.1.3.5 Otros desarrollos e implementaciones del HMC – Otros desarrollos como NUTS (= No U Turn Sampler, Homan & Gelman, 2014) se centran en las soluciones de los problemas prácticos de ajuste fino del algoritmo HMC – el tamaño del paso de los leapfrogs y el número de pasos de leapfrog repetidos *L* antes de hacer una propuesta. El nombre NUTS apunta al problema de evitar una curva en U invertida, es decir demasiados pasos de salto antes de hacer una propuesta y socavar así la optimización. El otro problema, el tamaño del escalón, también lo discuten los autores. El algoritmo asociado *dual averaging* aborda una solución adaptada y automática. En resumen, el NUTS optimiza *L* y ϵ tal que

- L – evita los giros en U
- ϵ – dual averaging optimización para llevar la tasa de aceptación hacia el óptimo.

La implementación más significativa del HMC en la variante No U Turn Sampler se encuentra actualmente en *Stan* (The Stan Development Team, 2019a), que es la implementación más prometedora del

HMC a la que se puede acceder desde varios lenguajes de programación. Stan está desarrollado por Andrew Gelman, Bob Carpenter y un equipo de desarrollo de más de 36 personas. En aplicación de la HMC, rápidamente se hace evidente que se pueden estimar modelos más complejos a través de Stan, que no necesariamente convergen con BUGS o JAGS o un algoritmo MH. Además, el tiempo hasta la convergencia es mensurablemente más corto. Sin embargo, esto requiere la recompilación del modelo estadístico en cuestión en lenguaje máquina, ya que de otro modo el tiempo necesario aumentaría enormemente. La compilación en sí no está relacionada con el algoritmo HMC como tal, sino que está en la naturaleza de los modelos complejos y sólo tiene algo que ver con la velocidad.

Lo que es importante cuando se utiliza NUTS y Stan es que no hay *burn-ins* visibles para eliminar más tarde. Todo lo que se emite corresponde al modelo posterior. Internamente, NUTS tiene una especie de fase de calentamiento para establecer los parámetros adecuados para el algoritmo. Todo lo que viene después son muestras reales de la distribución apuntada. Por lo tanto, si la cadena MCMC no se ve bien, como en el caso de la falta de convergencia, esto afecta a toda la cadena MCMC y no sólo a los primeros sorteos e indica problemas con el modelo, la Prior o los parámetros iniciales.

Una gran ventaja del algoritmo HMC es que, cuando falla, lo hace correctamente debido a unos valores a priori deficientes o a unos supuestos de distribución incorrectamente elegidos. El fallo aparece de forma tan evidente que hay que cambiar el modelo. McElreath (2015) da ejemplos aquí de lo que sale cuando se toma una Prior uniforme en lugar de una Prior ligeramente informada – no hay convergencia y no hay modelo razonable. Esto debería poner un poco de freno a los barridos a priori que no son sustantivos, como es el caso de la distribución uniforme. Es mejor proporcionar alguna dirección y, si es posible, introducir la regularización de este modo, para permitir valores posibles y excluir valores imposibles.

6.13.1.4 Resumen de los algoritmos MCMC

La ventaja del HMC sobre el algoritmo MH es que funciona de forma muy fiable en espacios de alta dimensión, incluso cuando el número de direcciones en el espacio de parámetros aumenta exponencialmente. Al mismo tiempo, sin embargo, generalmente sólo un pequeño número de posibilidades resultan ser típicas para la distribución de probabilidad apuntada (Betancourt, 2017. Con el HMC, el espacio de parámetros no se recorre simplemente al azar (por random walk), como en la MH, pero se recorre a saltos según las especificaciones de la dinámica hamiltoniana, de modo que se pueden generar más rápidamente muestras independientes entre sí. De este modo, en caso de problemas complejos se obtienen tasas de aceptación óptimas de hasta el 65 % para el HMC y, al mismo tiempo, sólo del 23 % para el MH.

Por otra parte, un problema del algoritmo HMC parece residir en la superación de mínimos locales aislados. Sin embargo, este problema también afecta al algoritmo MH y a otras variantes. La entrada del blog de Rogozhnikov (2016, véanse las figuras 6.81 y 6.82), `ptII_Bayes_quan_problem-local-minima.r`) es interesante porque proporciona una visualización gráfica interactiva del algoritmo MH frente a la simulación HMC, de modo que la tipicidad del algoritmo MH y del algoritmo HMC se pone de manifiesto de forma vívida. Al poder determinar tanto la dirección como la fuerza de impacto del movimiento de la HMC es posible que el algoritmo HMC supere en absoluto los mínimos locales. De este modo, los valles locales (mínimos locales) pueden superarse mucho mejor, o incluso en absoluto, en comparación con el algoritmo MH y, sobre todo, progresa más rápidamente. La estructura del algoritmo MH y la comparación con un número aleatorio distribuido uniformemente para aceptar o rechazar los candidatos de la cadena MCMC se mantiene tal como se ha descrito.

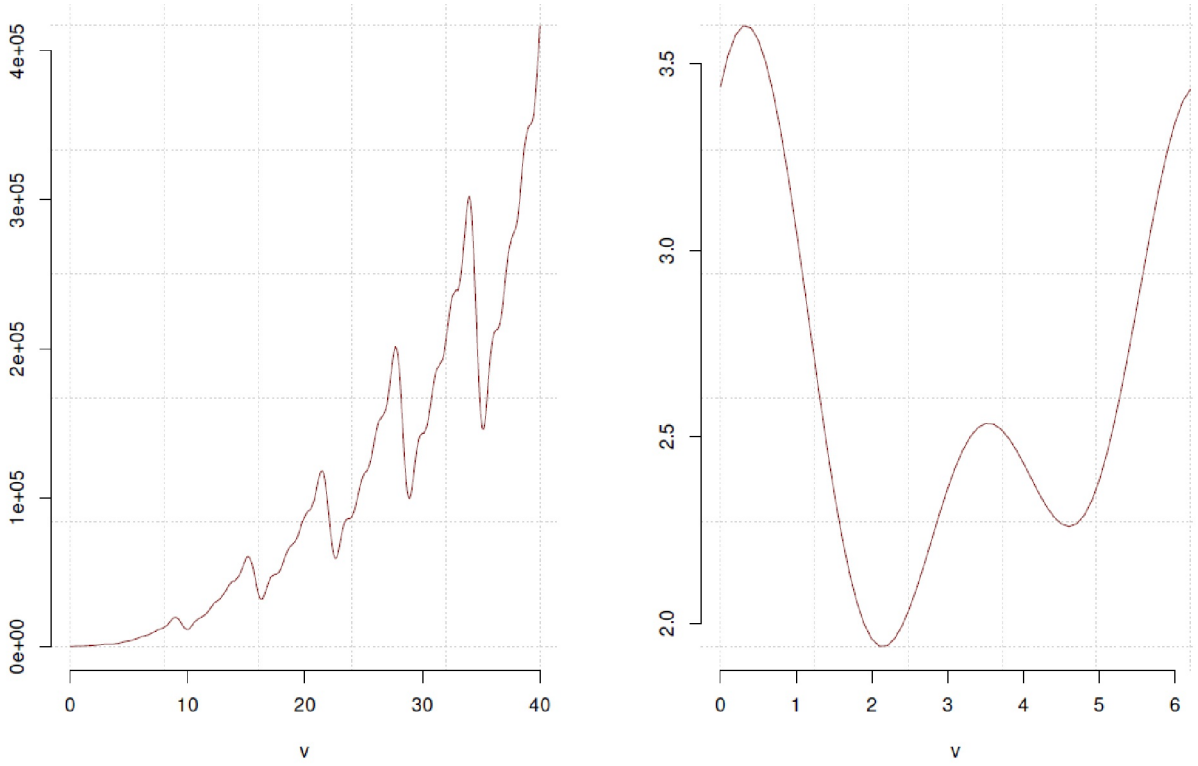


Figura 6.81. *MCMC – Problema de mínimos locales*

Si se observan los algoritmos MCMC desde un punto de vista físico (Neal, 2011), es decir, como un sistema con determinados estados energéticos, la diferencia entre los algoritmos resulta intuitivamente más obvia. El algoritmo MH simplemente toma el vecino del estado actual y examina el cambio de energía y acepta o rechaza este nuevo candidato. Como se recorre el espacio de parámetros a ciegas y sin pensar, el procedimiento es seguro, pero no muy eficiente, sobre todo en el rango de las altas dimensiones. Y cuando prevalecen los mínimos locales y otras condiciones difíciles, se convierte en un reto. La situación es aún más difícil cuando hay muchos parámetros que estimar, por ejemplo en los HLM. Entonces, un algoritmo MH puede alcanzar rápidamente sus límites para encontrar los pocos picos en el rango de alta dimensión. En cambio, el algoritmo HMC funciona dando al estado actual un empujón arbitrario (= energía cinética) y esperando a que se mueva un poco (= paso de leapfrog) para luego detener el sistema y calcular posteriormente la probabilidad de aceptación.

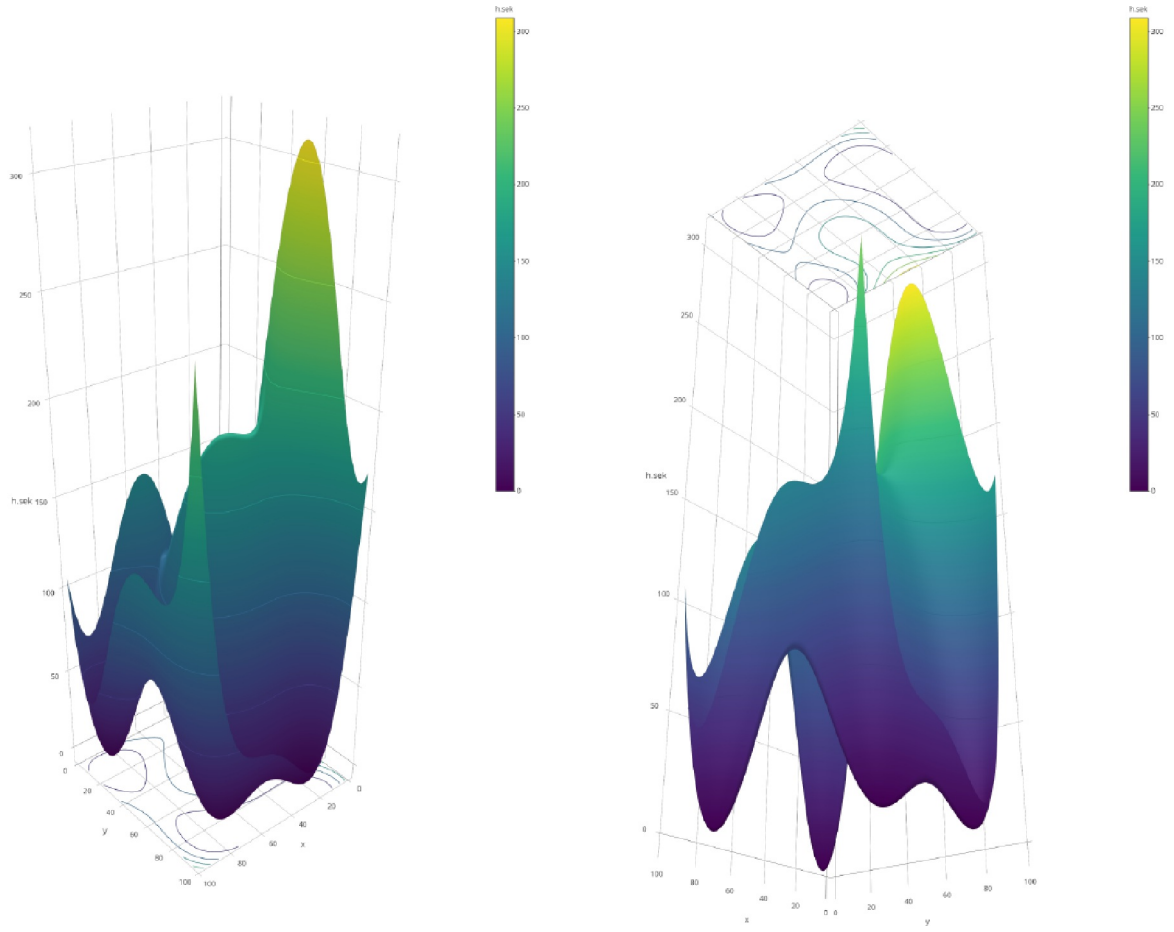


Figura 6.81. *MCMC – Problema de mínimos locales (Función de Himmelblau)*

De este modo, la posición y el impulso pueden modelizarse por separado, pero dan lugar a una distribución de probabilidad común. Una excelente animación de este proceso y el problema de la superación de mínimos locales se puede encontrar en Rogozhnikov (2016). Se puede imaginar la superación de los mínimos locales como un aumento sucesivo de la energía de las colisiones cinéticas, que obtienen así un mayor impulso, de modo que con el tiempo la probabilidad de superar mínimos locales se hace mayor una vez que se ha investigado exhaustivamente la zona local. Sin embargo, como las energías al principio y al final de las simulaciones MCMC son entonces claramente diferentes entre sí, es necesario un ajuste cuidadoso de los criterios de aceptación tomados del algoritmo MH. La superación de un mínimo local debería entonces ir acompañada de una disminución de los choques de energía cinética. Coordinar esto es un proceso extremadamente complejo, que puede probar usted mismo en la animación para hacerse una idea del problema básico predominante.

En R (CRAN, 2019a), se puede encontrar el algoritmo Metropolis-Hastings en los paquetes de R `Bolstad2`, `MHadaptive`, `EMC`, `mcmc`, `FME` o `ramcmc`, entre otros. Para el muestreo de Gibbs, existen los paquetes de R `MCMCpack`, `gibbs.met` o `lda`. Si el muestreo de Gibbs se realiza fuera de R utilizando `BUGS` o `JAGS`, pero controlado desde R, se utilizan los paquetes `BRugs`, `rbugs`, `R2WinBUGS`, `glmmBUGS`, `R2OpenBUGS` y `tsbugs` o `BayesianFirstAid`, `BEST`, `bfw`, `rjags`, `R2jags` o `runjags`. El algoritmo HMC puede utilizarse a través de los paquetes de R `rstan` o `brms`. Cabe señalar que no todos estos paquetes R explotan necesariamente todo el espectro de posibilidades de los programas MCMC externos, sino que se limitan a determinados análisis. Se puede analizar los resultados con los paquetes R `boa`, `bayesplot`, `coda` o `MCMCvis`. Además, existen muchos blogposts con código R que utilizan el algoritmo MH (por ejemplo,

Wilkinson, 2004b, 2010; Hartig, 2010; Chivers, 2012a, 2012b) o el muestreador de Gibbs (p.ej. Das, 2014; Wilkinson, 2004a; Oganisian, 2007) o el algoritmo HMC (Clark, 2016) en ejemplos relativamente sencillos y comprensibles. Joseph (2013) muestra con un ejemplo sencillo cómo visualizar en R el algoritmo MH de forma que se pueda observar directamente la aparición de la cadena MCMC. El paquete R `rethinking` permite hacer lo mismo para el HMC almacenando las trayectorias y haciéndolas accesibles de forma gráfica. De este modo, con la ayuda del paquete `HMCdirect` se puede seguir visualmente la aparición de una cadena MCMC

6.13.2 Ejemplos de algoritmos MCMC en R

6.13.2.1 Metrópolis-Hastings en R

El algoritmo MH y el muestreo de Gibbs pueden simularse con bastante facilidad en R, como muestra el siguiente ejemplo. Comenzamos con el algoritmo MH y la simulación simple de una distribución normal basada en datos empíricos. Se elige el caso de *media desconocida ⇆ varianza conocida*.

Dado que la distribución normal es su Prior conjugado, se pueden utilizar los parámetros μ_{post} y τ^2_{post} para la distribución posterior, esta es la varianza inversa y se utiliza a menudo en los programas MCMC como medida de la precisión y, por tanto, de la varianza (`ptII_quan_Bayes_Metropo-lisHastings_ejemplo-normdist.r`). A continuación se determinan la varianza de la Likelihood y de la Prior, la media de la Prior y el tamaño de la muestra *n*. Éstos podemos tomar para generar *n* datos distribuidos normalmente con $\mu =$ ... y $\sigma =$ ... respectivamente podríamos tomar un conjunto de datos existente como los datos de Darwin (conjunto de datos `darwin.maize` del paquete R `agridat`, s. cap. 6.8.4.2.1).

```
# or use Darwin's data R-Code
y <- agridat::darwin.maize[,"height"]
y
# summary
as.data.frame(t(c(N=length(y),summary(y),SD=sd(y),VAR=var(y),
fivenum2(y))))
# analytical solution
n <- length(y)
# mu prior
mu <- 20
# sample size
# n <- 30
# variance likelihood
s2 <- sd(y)
# variance prior tau^2
t2 <- 1/1e2
# variance of random draw = candidate = thetaST(ar)
s2.prop <- 2
```

Ahora podríamos ver la solución analítica, ya se puede actualizar de modo bayesiano la distribución normal vía conjugación:

```
# normal distribution is a conjugate prior
# calculate theoretical true posterior parameters
# mean of the normal posterior
mu.n <- ( mean(y)*n/s2 + mu/t2 ) / ( n/s2+1/t2)
# precision of the normal posterior
t2.n <- (n/s2+1/t2)
t2.n
# mu posterior
```

```
mu.n
# variance posterior
s2.n <- 1/t2.n
s2.n
# sd posterior
sqrt(s2.n)
```

Como alternativa determinamos unos pocos valores de entrada para la simulación MH o tamizamos los existentes, incluyendo el número de simulaciones, el valor inicial del parámetro θ de interés y la varianza de las extracciones aleatorias s^2.

```
> # number simulations
> nsim <- 1e+5
> # initial value of theta (= parameter of interest)
> theta0 <- 20
> s2
[1] 3.180953
> t2
[1] 0.01
> s2.prop
[1] 2
```

Con estos valores se llama a la función de R `MH_norm()`.

```
# seed R-Code
seed <- 1889
# run MH
mat <- MH_norm(y=y, s2.prop=s2.prop, s2=s2, t2=t2, mu=mu,
        theta0=theta0, nsim=nsim, seed=seed)
theta.post <- mat[,"theta"]
```

Éstos son los primeros y los últimos valores de la tabla:

```
> head(mat)
     thetaST   u.log      r.log    theta acceptance
[1,] NA        NA         NA       20     NA
[2,] 23.47224 -0.9385766 -696.2426 20     0
[3,] 22.61829 -0.2506334 -402.6739 20     0
[4,] 18.83300 -0.3764782  -62.2267 20     0
[5,] 21.67170 -0.3184328 -170.5125 20     0
[6,] 18.43309 -1.4710669 -117.8356 20     0
> tail(mat)
           thetaST   u.log     r.log     theta    acceptance
[99995,]  18.47740 -0.7409499 -110.58157 20.01982 0
[99996,]  18.55703 -2.4519040  -98.50012 20.01982 0
[99997,]  22.55093 -0.3285948 -382.68081 20.01982 0
[99998,]  21.54463 -0.8850934 -146.58104 20.01982 0
[99999,]  18.84070 -1.0752333  -61.09730 20.01982 0
[100000,] 19.49506 -1.0262288   -8.40233 20.01982 0
```

Primero eliminamos la secuencia Burn-in (= quemarse). En nuestro caso, se trata de las 500 primeras simulaciones MCMC, que sirven para estabilizar la secuencia:

```
# remove first few estimations from theta.post
theta.post.noburnin <- theta.post[-c(1:500)]
```

Como salida obtenemos la cadena MCMC, para la que podemos obtener estadísticas resumidas

```
> # summary
> list(summary=summary(theta.post),sd=sd(theta.post),
+ var=var(theta.post),fivenum=fivenum2(theta.post))
$summary
Min.  1st Qu. Median Mean  3rd Qu. Max.
19.56 19.84   19.90  19.90 19.97   20.28
$sd
[1] 0.09600736
$var
[1] 0.009217413
$fivenum
minimum  lower-hinge median   upper-hinge maximum
19.55798 19.83720    19.90314 19.96776    20.28370
```

y trazar con `plotPost()` desde BEST (véase Fig. 6.83):

```
# plot posterior
compVal <- 20
par(oma=c(2,1,2,1), "cex.axis"=1, bty="l")
plotPost(theta.post, credMass=0.87, compVal=compVal,
       ROPE=c(119.6,20.2), xlim=c(19.5,20.2), xlab=expression(theta))
lines(density(theta.post), col="darkred", lwd=2, lty=2)
mtext("MC-Simulation via Metropolis Hastings",
       outer=TRUE, line=-1, cex=1.5, side=3)
mtext(expression(paste("Normal values for unknown parameter ",
       theta,"",sep="")), outer=TRUE, line=-2.5, cex=1, side=3)
```

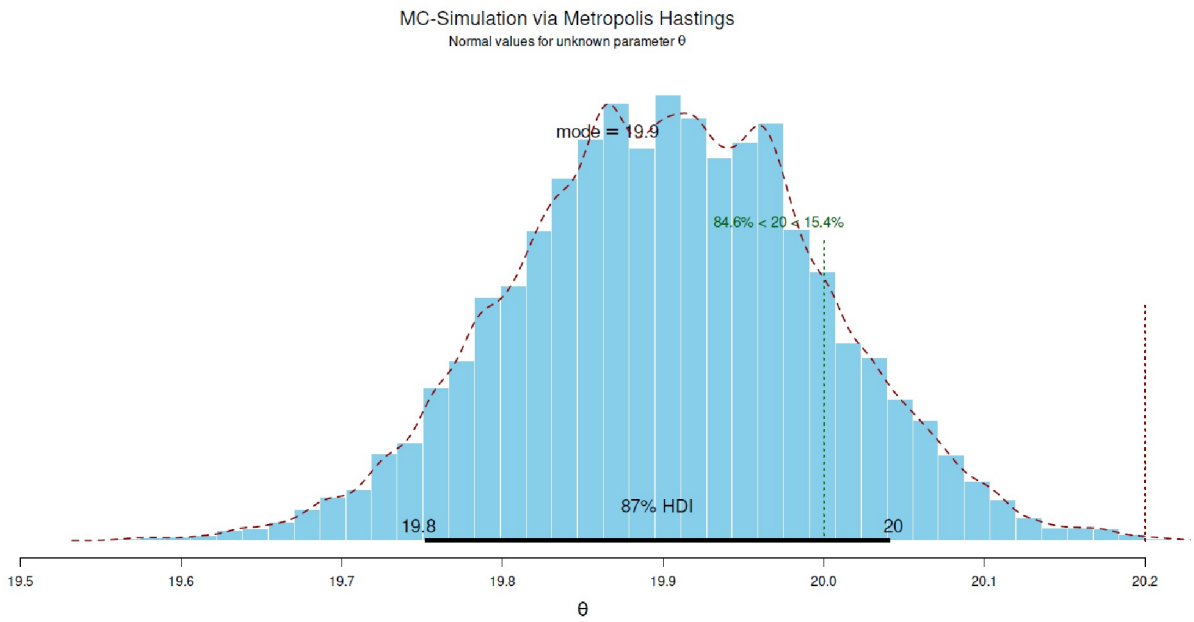


Figura 6.83. *Algoritmo MH (Posterior)*

Con `plot()` y otras funciones de `coda`, se puede mostrar el curso de la cadena MCMC (véase la Fig. 6.84) o HDI:

```
# MCMC chain diagnostics
theta.mcmc <- as.mcmc(theta.post)
```

```
plot.mcmc(theta.mcmc, col="darkred")
par(mfrow=c(2,2))
densplot(theta.mcmc, col="darkred", bty="n")
traceplot(theta.mcmc, bty="n", col="darkred")
#autocorr.plot(theta.mcmc)
acf(theta.mcmc)
summary(as.mcmc(theta.post))
hdi(theta.post, credMass=0.87)
> hdi(theta.post, credMass=0.87)
lower upper
19.75199 20.04079
attr(,"credMass")
[1] 0.87
```

Esto se puede comparar con la misma cadena si, por ejemplo, se eliminan las 500 primeras extracciones ("burn in").

```
> # compare with sequence without burn-ins
> # no real difference
> plot.mcmc(as.mcmc(theta.post.noburnin), col="darkred")
> list(summary=summary(theta.post.noburnin),
       + sd=sd(theta.post.noburnin),
       + var=var(theta.post.noburnin),
       + fivenum=fivenum2(theta.post.noburnin))
$summary
Min.  1st Qu. Median Mean  3rd Qu. Max.
19.56 19.84   19.90  19.90 19.97   20.28
$sd
[1] 0.09596093
$var
[1] 0.009208501
$fivenum
minimum  lower-hinge median   upper-hinge maximum
19.55798 19.83710    19.90314 19.96776    20.28370
```

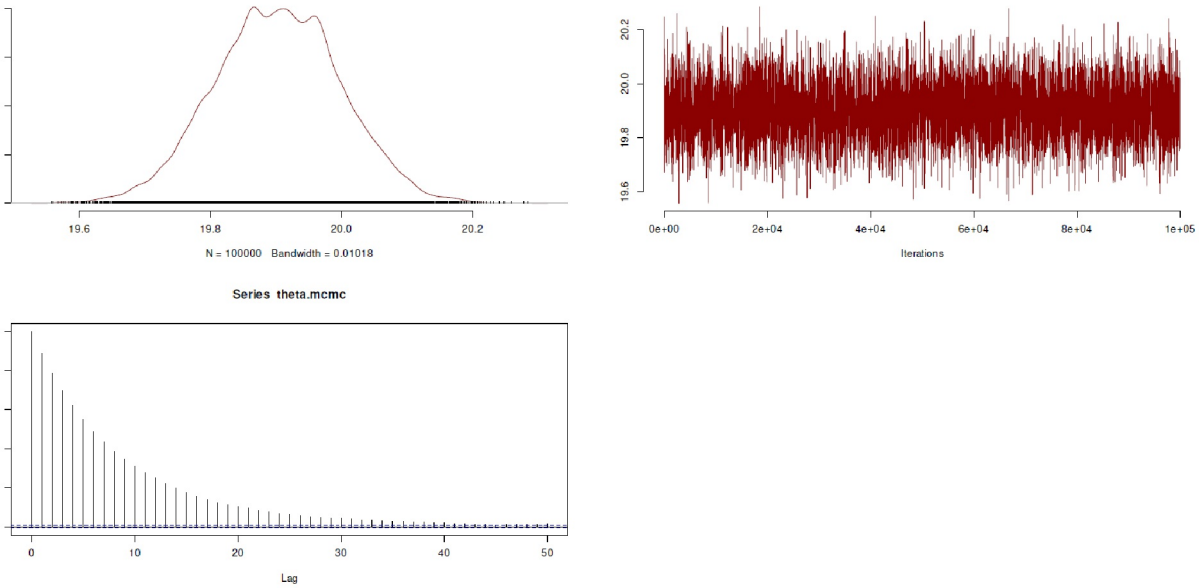


Figura 6.84. *Algoritmo MH (Diagnóstico MCMC)*

Las tasas de aceptación o rechazo son igualmente interesantes, numéricamente

```
> mat.red <- mat[-1,]
> # acceptance rate
> acc.r <- sum(mat.red[,"acceptance"])/nsim
> # rejection rate
> rej.r <- 1-acc.r
> acc.r
[1] 0.08648
> rej.r
[1] 0.91352
>
> # acceptance versus rejection rate
> acc.r/rej.r
[1] 0.09466678
> 1-(acc.r/rej.r)
[1] 0.9053332
> acc.r.cs <- cumsum(mat.red[,"acceptance"])
> head(acc.r.cs)
[1] 0 0 0 0 0 0
> tail(acc.r.cs)
[1] 8648 8648 8648 8648 8648 8648
```

y como gráfico (véase Fig. 6.85):

```
# plot acceptance rate R-Code
par(oma=c(2,1,2,1), "cex.axis"=1, bty="l", mfrow=c(1,2))
plot(acc.r.cs, type="l", xlab="draw", ylab="acceptance rate (cumsum)",
       col="violetred3", pre.plot=grid(), bty="n")
# looks like a straight line, zoom in:
plot(acc.r.cs[1:100], type="l", xlab="draw", ylab="acceptance rate (cumsum)",
       col="violetred3", pre.plot=grid(), bty="n")
mtext("Metropolis Hastings MCMC", outer=TRUE, line=-1, cex=1.5, side=3)
mtext(expression(paste("Cumulated acceptance rates (zoom in)", sep="")),
       outer=TRUE, line=-2.5, cex=1, side=3)
```

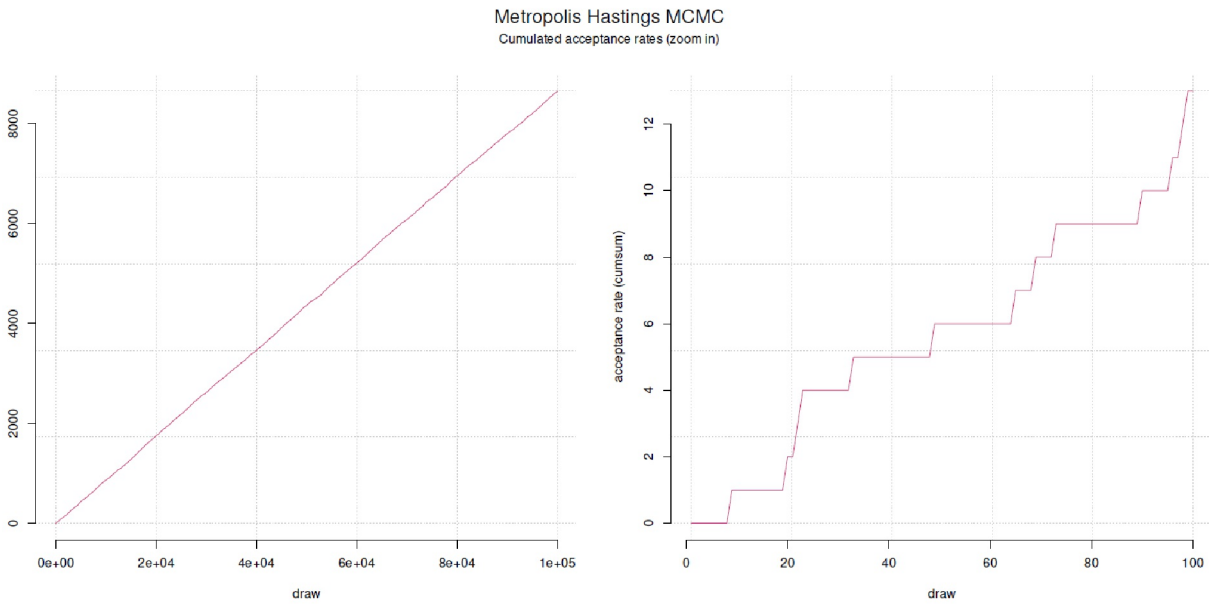


Figura 6.85. *Algoritmo MH (tasa de aceptación)*

Del mismo modo, se pueden comparar los valores θ aceptados con los valores θ^* posibles (candidatos), véase Fig. 6.86 (completa y como extracto para el primer centenar de los candidatos). Más análisis son concebibles y útiles (por ejemplo, `gelman.plot()` de `coda`).

```
# plot thetaST(ar) vs theta R-Code
# = proposed candidate vs. chosen candidate
par(oma=c(2,1,2,1), "cex.axis"=1, bty="l")
plot(mat[,"thetaST"],mat[,"theta"], pch=21, cex=1, bg="gray10",
       col="violetred3", bty="n", pre.plot=grid(), ylab=expression(theta),
       xlab=eval(substitute(expression(paste(theta,"*",sep="")))),
       main="", cex.lab=1.2)
# just an outtake
START <- 1
END <- 100
plot(mat[,"thetaST"][START:END],mat[,"theta"][START:END],
       type="b", col="darkred", pre.plot=grid(), bty="n", ylab=expression(theta),
       xlab=eval(substitute(expression(paste(theta,"*",sep="")))),
       main="", cex.lab=1.2)
mtext("Metropolis Hastings MCMC", outer=TRUE, line=-1, cex=1.5, side=3)
mtext(eval(substitute(expression(paste(theta,"* (= candidate) versus ",
       theta," (= chosen) for full and ",START," to ",END," steps",
       sep="")),list(START=START,END=END))), outer=TRUE, line=-2.5, cex=1, side=3)
```

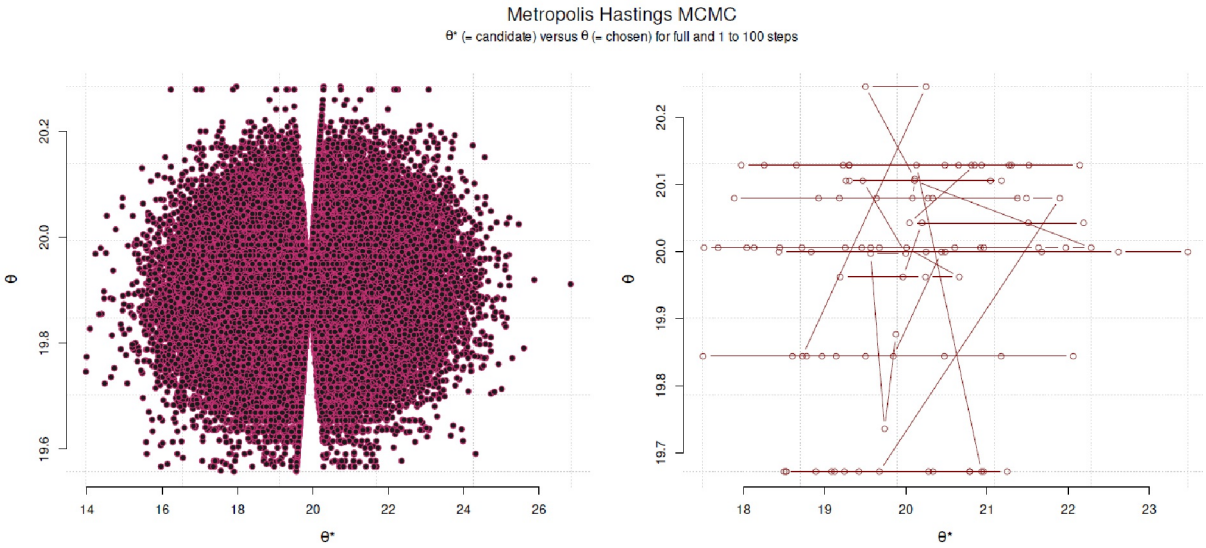


Figura 6.86. *Algoritmo MH (valores aceptados frente a candidatos, completo y extracción)*

6.13.2.2 Muestreo de Gibbs en R

Para el muestreo de Gibbs, primero recurrimos a la función de R `mu.sigma2.post()`. Aplicamos datos empíricos, aquí un conjunto de datos de Hoff (2009, cap. 10) con las estadísticas descriptivas estadísticos, para los que se desconoce la media μ de la población y se conoce la desviación estándar σ de la población (`ptII_quan_Bayes_GibbsSampling_example-normdist.r`).

```
# From: A First Course in Bayesian Statistical Methods
# by Peter D. Hoff (2009-07-14), chap. 10
y <- c(9.37, 10.18, 9.16, 11.60, 10.33)
y
```

```
# sample size
n.emp <- length(y)
# empirical summary
list(N=n.emp,summary=summary(y),sd=sd(y),var=var(y))
```

Como valores a priori eligimos $\mu_{prior} = 5$, $\sigma^{2prior} = 10$ y $\sigma = 1$.

```
# prior values R-Code
mu.prior <- 5
sigma2.prior <- 10
mu <- 10
sigma <- 1
sigma2 <- sigma^2
tau2 <- 1/sigma2
```

La función de R `mu.sigma2.post()` toma μ_{prior}, σ^{2prior} y σ y determina analíticamente mediante la fórmula de la conjugación para este caso de distribución normal μ_{post} y σ_{post}. Si no se dispusiera de datos empíricos, se podrían utilizar los valores a priori y una desviación estándar para los valores de la población con `rnorm()`.

```
> # create posterior values via conjugation
> # from prior and likelihood
> # mu.sigma2.res <- mu.sigma2.post(y=c(9.37,10.18,9.16,11.60,10.33),
        + mu.prior=5, sigma2.prior=10,
        + sigma2.data=1)
> mu.sigma2.res <- mu.sigma2.post(y=y, mu.prior=mu.prior,
        + sigma2.prior=sigma2.prior,
        + sigma2.pop=sigma2)
> mu.sigma2.res
mu.prior sigma2.prior sigma2.pop mu.post  s2.post   s.post    tau
5        10           1          10.02745 0.1960784 0.4428074 5.1
```

Con los valores a posteriori muestreamos mediante `rnorm()` a partir de la Posterior – el muestreo de Gibbs real. En el muestreo de Gibbs no existe una consulta de aceptación como en el algoritmo Metropolis-Hastings.

```
# create "draw" values from posterior
nsamps <- 1e5
mc.res <- rnorm(n=nsamps, mean=mu.sigma2.res[,"mu.post"], sd=mu.sigma2.res[,"s.post"])
```

Trazamos y analizamos los valores posteriores generados de este modo como antes (véase la Fig. 6.87)

```
# plot
par(oma=c(2,1,2,1), "cex.axis"=1, bty="l")
plotPost(mc.res, credMass=0.87, compVal=10, ROPE=c(9.5,10.5),
        showMode=TRUE, xlab=expression(mu), ylab="Density", main="")
lines(density(mc.res), lty=2, lwd=2, col="darkred")
mtext("MC-Simulation from posterior", outer=TRUE, line=-1.2, cex=1.5, side=3)
mtext("Normal values derived from analytical posterior",
        outer=TRUE, line=-2.5, cex=1, side=3)
```

y compararlos con la salida de `normnp()` del paquete `Bolstad` de R. Los valores posteriores para μ y σ coinciden exactamente, lo que no es sorprendente en vista de las fórmulas disponibles para la conjugación.

```
> # compare with function from R package 'Bolstad'
> # check and compare results
```

```
> mu.sigma2.res
mu.prior sigma2.prior sigma2.pop mu.post  s2.post   s.post    tau
1        5            10         10.02745 0.1960784 0.4428074 5.1
> list(summary=summary(mc.res),sd=sd(mc.res),var=var(mc.res))
$summary
Min.  1st Qu. Median Mean   3rd Qu. Max.
8.159 9.728   10.027 10.026 10.324  11.966
$sd
[1] 0.4434986
$var
[1] 0.196691
> summary(as.mcmc(mc.res))
Iterations = 1:1e+05
Thinning interval = 1
Number of chains = 1
Sample size per chain = 1e+05
1. Empirical mean and standard deviation for each variable,
        plus standard error of the mean:
Mean       SD       Naive SE  Time-series SE
10.026152 0.443499 0.001402  0.001402
2. Quantiles for each variable:
2.5%  25%   50%    75%    97.5%
9.160 9.728 10.027 10.324 10.894
> # compare with Bolstad package
> normnp(x=y, m.x=mu.prior, s.x=sqrt(sigma2.prior), sigma.x=1, mu=NULL, plot=FALSE)
Known standard deviation :1
Posterior mean : 10.027451
Posterior std. deviation : 0.4428074

Prob. Quantile
------ ----------
0.005 8.8868546
0.010 8.9973268
0.025 9.1595643
0.050 9.2990976
0.500 10.0274510
0.950 10.7558044
0.975 10.8953376
0.990 11.0575751
0.995 11.1680474
```

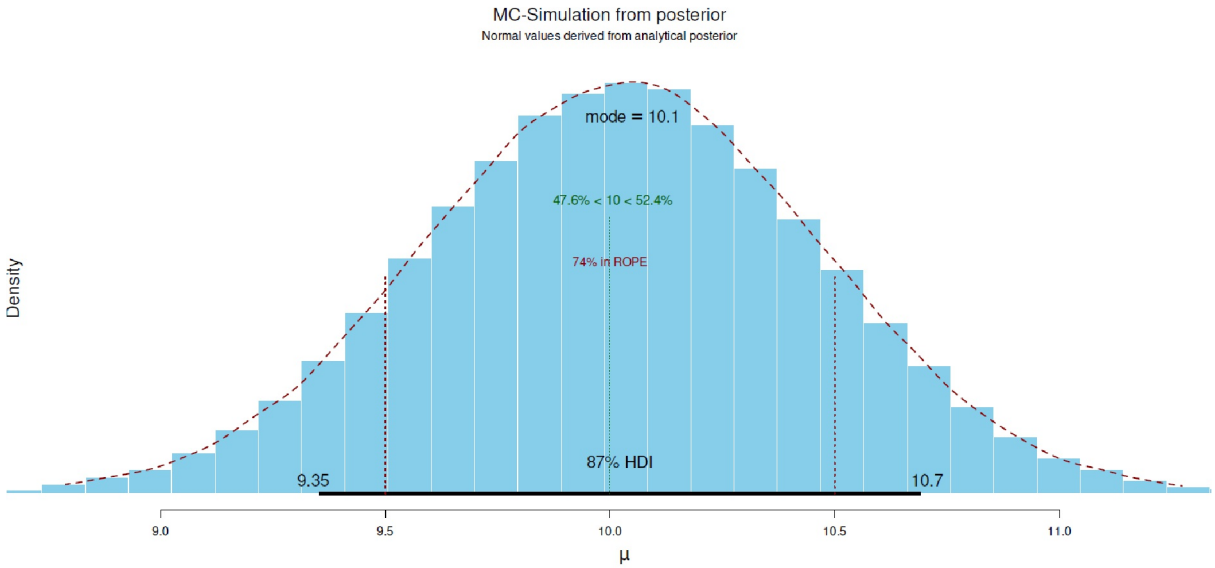


Figura 6.87. *Muestreo de Gibbs (Posterior)*

6.13.2.2.1 Muestreo de Gibbs con JAGS – Ahora repetimos lo mismo con JAGS. En primer lugar se define el modelo JAGS (ptII_quan_Bayes_JAGS_example-norm.r).

```
# define model for JAGS
model_string <- "model{
# Likelihood
# inv.var = tau = 1/var
for(i in 1:n){
Y[i] ~ dnorm(mu,tau)
}
# Prior for mu
mu ~ dnorm(mu.prior, tau0)
tau0 <- 1/sigma2.prior
# Prior for the inverse variance
tau ~ dgamma(a, b)
# Compute the variance
sigma2 <- 1/tau
}
"
# end of JAGS model
```

y, a continuacion, lo comprobamos

```
# check
cat(model_string)
```

y luego tomamos los valores a priori o los datos empíricos del ejemplo de muestreo de Gibbs anterior. El parámetro de precisión τ se alimenta de una distribución T con los parámetros *a* y *b*, para los que elegimos $a = 2$ y $b = 2$. Si fijáramos la precisión de forma diferente, digamos $a = 0{,}01$ y $b = 0{,}01$, los resultados serían significativamente diferentes debido a esta expectativa previa y las estimaciones serían más inciertas, especialmente para σ. Los lectores interesados pueden comprobarlo por sí mismos en R.

```
seed <- 1999 R-Code
set.seed(seed)
dig <- 3
# sample size
n.emp <- length(y)
# prior values
mu.prior <- 5
sigma2.prior <- 10
# dgamma prior parameters for tau
a <- 2
b <- 2
sigma <- 1
```

JAGS se llama con jags.model() y update() desde rjags y las cadenas MCMC se preparan con coda.samples().

```
> # bring model into JAGS
> model <- jags.model(textConnection(model_string), n.chains=3,
        + data=list(Y=y, n=n.emp, mu.prior=mu.prior,
        + sigma2.prior=sigma2.prior, a=a, b=b))
Compiling model graph
Resolving undeclared variables
Allocating nodes
Graph information:
Observed stochastic nodes: 5
Unobserved stochastic nodes: 2
Total graph size: 15
```

```
Initializing model
> # number of burn-ins for 10000 samples
> burnin <- 1e3
> n.iter <- 1e5
> # run model with JAGS
> update(model, burnin, n.iter=n.iter, progress.bar="none");
```

De la Posterior sacamos con `coda.samples()` y miramos las simulaciones MCMC.

```
# number of iterations for MCMC samples
iterats <- 1e4
# create samples from posterior
samps <- coda.samples(model, variable.names=c("mu","sigma2","tau"),
        n.iter=iterats, progress.bar="none")
str(samps)
head(samps[[1]])
tail(samps[[1]])
```

La Posterior puede ser elegantemente trazado con `plotPost()` de BEST o examinado con otros métodos de paquetes R como `coda`. Esto va seguido de una comparación con los valores determinados analíticamente para la Posterior y los gráficos para la autocorrelación, el tamaño efectivo de la muestra y el diagrama de Gelman con los diagnósticos de Gelman (Brooks & Gelman, 1998).

```
> # various MCMC plots
> plot(samps)
> autocorr.plot(samps)
> # sample size adjusted for autocorrelation
> eS <- effectiveSize(samps)
> eS
mu        sigma2   tau
28036.92 22941.78 25228.30
> ratio <- eS/ (length(samps)* dim(samps[[1]])[1])
> 1-ratio
mu         sigma2    tau
0.0654360 0.2352741 0.1590568
> # evolution of Gelman and Rubin's shrink factor
> # as the number of iterations increases
> gelman.plot(samps)
> gelman.diag(samps)
Potential scale reduction factors:
        Point est. Upper C.I.
mu      1          1
sigma2  1          1
tau     1          1
Multivariate psrf
1
```

Los gráficos dicen algo sobre si las extracciones MCMC de los valores posteriores dependen unas de otras, lo que no deberían. El tamaño de muestra efectivo ajusta el tamaño de muestra para las autocorrelaciones a lo largo del tiempo, y cuanto más se acerque este valor al tamaño real de la muestra (aquí: cadenas MCMC en la Fig. 6.88), lo mejor. Los diagnósticos de Gelman o el diagrama de Gelman por encima o lejos de 1 indican una falta de convergencia de las cadenas MCMC (véase la Fig. 6.89).

Dado el pequeño tamaño de la muestra de $N = 5$, todos estos valores están en orden y no dan ninguna indicación de que las cadenas MCMC no hayan convergido. Sin embargo, es obvio que la estimación de σ es bastante inexacta, mientras que la de μ parece mucho más precisa. Comparemos JAGS con `normnp()`:

```
> summary(samps)
Iterations = 100001:110000
Thinning interval = 1
Number of chains = 3
Sample size per chain = 10000
1. Empirical mean and standard deviation for each variable,
plus standard error of the mean:
        Mean  SD     Naive SE Time-series SE
mu      9.997 0.5038 0.002909 0.003011
sigma2  1.290 0.8922 0.005151 0.005890
tau     1.034 0.5169 0.002985 0.003255
2. Quantiles for each variable:
       2.5%   25%    50%    75%    97.5%
mu     8.9545 9.6958 10.016 10.320 10.941
sigma2 0.4434 0.7553  1.052  1.532  3.582
tau    0.2792 0.6529  0.951  1.324  2.255
> # compare with Bolstad package
> normnp(x=y, m.x=mu.prior, s.x=sqrt(sigma2.prior), sigma.x=NULL,
+ mu=NULL, plot=FALSE)
Standard deviation of the residuals :0.9646
Posterior mean : 10.0343144
Posterior std. deviation : 0.4274274
Prob. Quantile
------ ----------
0.005 8.9333344
0.010 9.0399696
0.025 9.1965721
0.050 9.3312589
0.500 10.0343144
0.950 10.7373699
0.975 10.8720567
0.990 11.0286592
0.995 11.1352944
> # =
> # normnp(x=y, m.x=mu.prior, s.x=sqrt(sigma2.prior), mu=mu, plot=FALSE)
```

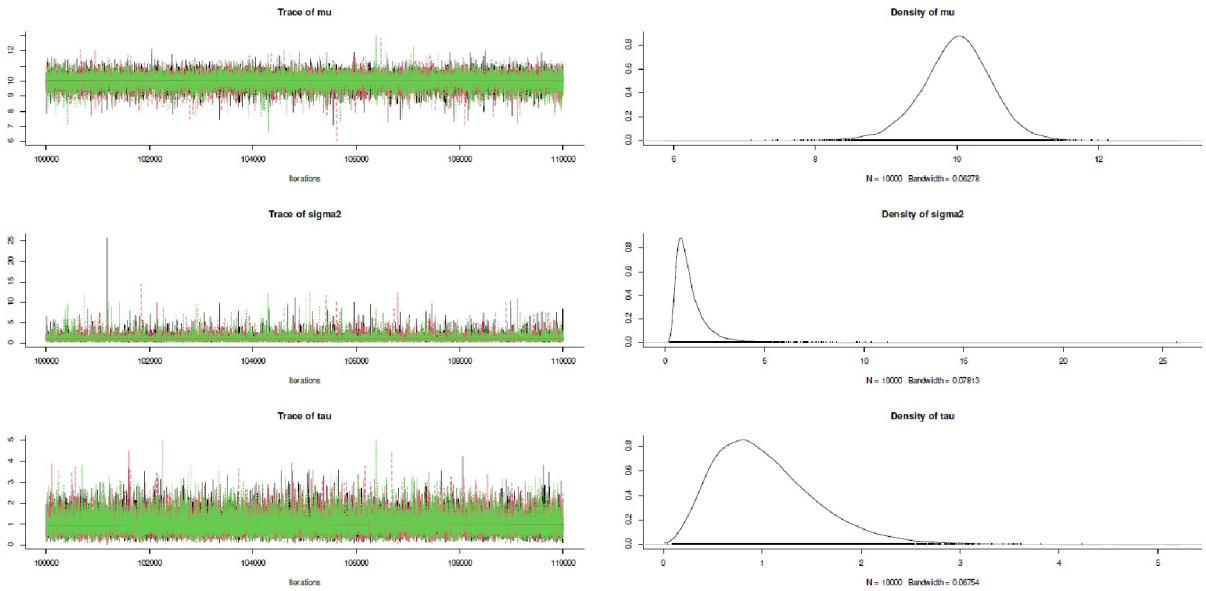


Figura 6.88. *Muestreo de Gibbs (diagnóstico MCMC, cursos)*

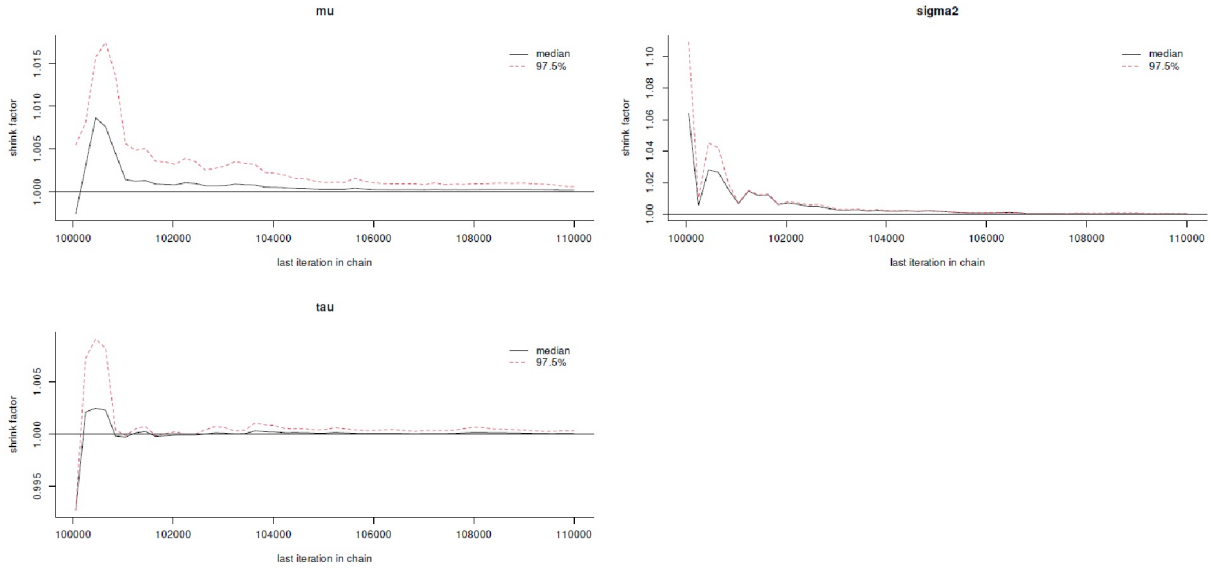


Figura 6.89. *Muestreo de Gibbs (diagnóstico MCMC, diagrama de Gelman)*

Una extensión, el caso de *media desconocida <–> varianza desconocida*, es posible en R con sólo unos pocos cambios. Partimos de los valores empíricos de una distribución – tamaño muestral, media y varianza, es decir, n, $\bar{x}$ y s^2 (`ptII_quan_Bayes_GibbsSampling_example-normdist.r`):

```
# another example, same procedure, different values
digs <- 3
seed <- 112
set.seed(seed)
# Gibbs sampling for mu and TAU (inverse variance)
# sample statistics
n <- 30
xbar.data <- 15
s2.data <- 3
```

Ahora definimos el tamaño de la muestra y la longitud de la secuencia de quemado. La longitud de quemado es las 1 000 primeras repeticiones. Si esto tiene sentido se puede comprobar más tarde con un `traceplot()` de `coda`.

```
n.burnin <- 1e3 R-code
sampsize <- 1e4 + n.burnin
```

Se toma la muestra de la distribución posterior conjunta de μ y τ, la varianza inversa.

```
# sample from the joint posterior(mu, TAU | data) R-Code
mu <- rep(NA, sampsize)
TAU <- rep(NA, sampsize)
# starting value for TAU (= intialization)
TAU[1] <- 1
for(i in 2:sampsize)
{
mu[i]  <- rnorm(n=1, mean=xbar.data, sd=sqrt(1 / (n*TAU[i-1])))
TAU[i] <- rgamma(n=1, shape=n/2, scale=2 /
          ((n-1) * s2.data + n * (mu[i]-xbar.data)^2))
}
```

Los parámetros μ (de distribución normal) y τ (de distribución gamma) se muestrean a partir de *n*, $\bar{x}$ y s^2 mediante `rnorm()` y `rgamma()` respectivamente. Como puede verse en las llamadas, las siguientes son relevantes para determinar el valor medio $\bar{x}$ y la desviación estándar *s*. Esta última se calcula a partir del tamaño de la muestra *n* y los datos disponibles hasta la simulación respectiva mediante la varianza inversa τ. La varianza inversa τ se utiliza a menudo como parámetro de precisión. Estos datos constituyen los parámetros de `rnorm()`. Para el cálculo de τ se sabe que la forma distribuida gamma o el parámetro de escala es una derivada del tamaño de la muestra, la varianza empírica y la desviación al cuadrado de los valores simulados hasta ahora para μ vs. la media empírica $\bar{x}$, que se calcula con el tamaño de la muestra *n*.

A continuación, se eliminan los burn-ins y se observan los resultados:

```
> # remove burnin
> mu  <- mu[-(1:n.burnin)]
> TAU <- TAU[-(1:n.burnin)]
> list(summary=summary(mu),sd=sd(mu),var=var(mu))
$summary
Min.  1st Qu. Median Mean  3rd Qu. Max.
13.71 14.78   15.00  15.00 15.21   16.64
$sd
[1] 0.3276221
$var
[1] 0.1073362
> list(summary=summary(TAU),sd=sd(TAU),var=var(TAU))
$summary
Min.    1st Qu. Median  Mean    3rd Qu. Max.
0.09217 0.27054 0.32524 0.33285 0.38762 0.73467
$sd
[1] 0.08651299
$var
[1] 0.007484498
```

y todo se traza (no se imprime):

```
# transform to MCMC object to process with functions from 'coda'
traceplot(as.mcmc(mu), col="darkred")
traceplot(as.mcmc(TAU), col="purple")
par(oma=c(2,1,2,1), "cex.axis"=1, bty="l", mfrow=c(1,2))
plotPost(mu, xlab=expression(mu),)
lines(density(mu), col="darkred", lwd=2, lty=2)
plotPost(TAU, xlab=expression(tau))
lines(density(TAU), col="darkred", lwd=2, lty=2)
mtext("MC-Simulation via Gibbs sampling", outer=TRUE, line=-1, cex=1.5, side=3)
mtext(expression(paste("Normal values for ",mu," (= mean) and ",tau,
       " (= inverse variance)",sep="")), outer=TRUE, line=-2.5, cex=1, side=3)
```

Una comparación de la correlación de μ y τ muestra que la correlación se disminuye con la distacia de las burn-ins:

```
> r1 <- cor(mu,TAU)
> r1
[1] 0.0126719
> # remove burnins (must be more in other cases)
> r2 <- cor(mu[-c(1:n.burnin)],TAU[-c(1:n.burnin)])
> r2
[1] 0.01247159
> r1/r2
[1] 1.016062
```

6.13.2.2 Otros gráficos para el muestreo de Gibbs

Las funciones de R `plot.mcmc()` y `plot.mcmc.parts()` muestran algunos de los gráficos mencionados. Relevante es la transferencia de μ estimado por MCMC; la función `plot.mcmc()` devuelve los histogramas, la estimación de la densidad, la cadena MCMC y la autocorrelación. Con `plot.mcmc()` se añade la posibilidad de trazar el desarrollo y el curso de la cadena MCMC para un determinado rango, por ejemplo, los 100 primeros o los comprendidos entre 2001 y 2999. Las cadenas MCMC para el rango seleccionado se imprimen en la parte superior, y las correlaciones de μ y τ así como el curso de las correlaciones a través de las simulaciones se imprimen debajo (ptII_quan_Bayes_GibbsSampling_example-normdist.r). Tomamos los resultados anteriores con las variables π y τ. Primero la simulación MCMC completa (véase la Fig. 6.90):

```
plot.mcmc(mu, TAU)
```

Ahora sólo trazaremos extractos que muestran de forma impresionante cómo se acumula lentamente una simulación MCMC de este tipo. Ahora vienen las primeras 20 simulaciones (ver Fig. 6.91 arriba) y luego las de 200 a 300 (véase la Fig. 6.91):

```
plot.mcmc.parts(mu,TAU, part=1:20)
plot.mcmc.parts(mu,TAU, part=200:300)
```

Por supuesto, podríamos extraer episodios posteriores (no impresos):

```
plot.mcmc.parts(mu,TAU, part=1:300) R-Code
plot.mcmc.parts(mu,TAU, part=1:nsim)
```

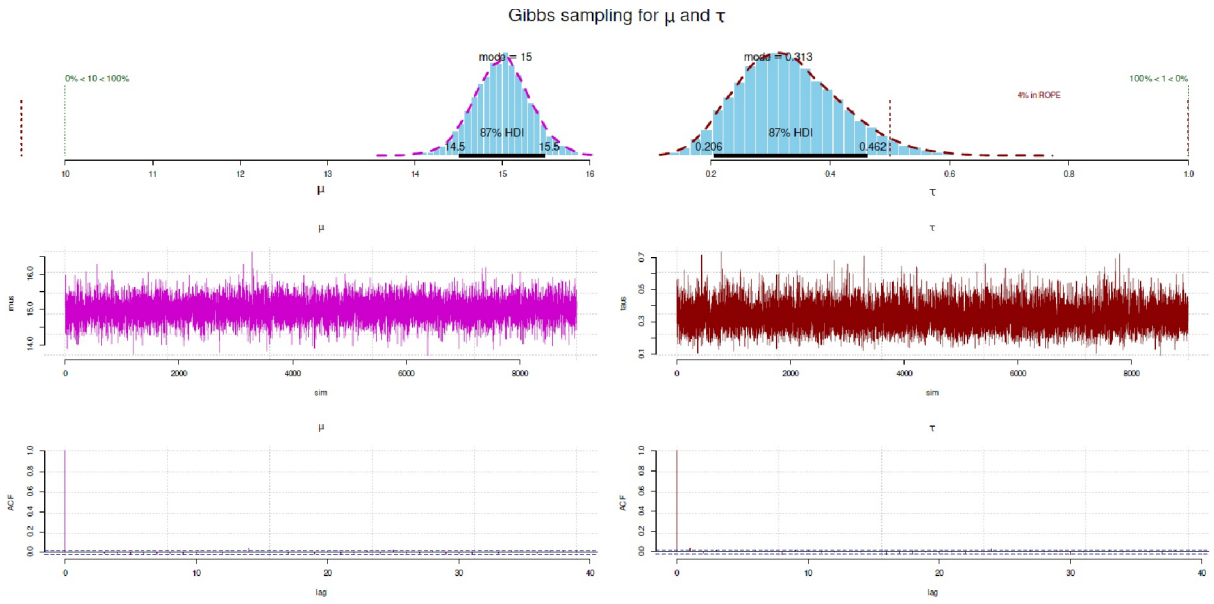


Figura 6.90. *Muestreo de Gibbs (secuencia MCMC completa)*

6.13.2.3 Hamilton Monte Carlo en la R

Recordemos que el muestreador HMC es una forma muy inteligente y sofisticada del algoritmo Metropolis-Hastings (véase el capítulo 6.13.1.1). Para la demostración en R, hacemos uso de un ejemplo elaborado del artículo básico de Neal (2011, p.125), al que hacen referencia muchos scripts de R. El paquete de R `hmclearn` se basa, además del artículo de Neal (ibíd.) y Betancourt (2017), en el trabajo de Thomas y Tu (2020) y proporciona más opciones de implementación para comprender mejor el proceso a nivel de código. El paquete R `rethinking` contiene una función R equivalente `HMC2()`, que toma directamente el script R de Neal y lo amplía con más resultados útiles para el diagnóstico sobre las trayectorias. Otros paquetes de R como `rhmc, hmc, johnnylaw` o `greta` contienen implementaciones más o menos comparables. En comparación con `rstan` y `brms`, sin embargo, son paquetes R más fáciles de usar, ya que con `brms` basta con utilizar la notación R habitual para modelos.

6.13.2.3.1 Código de modelo R según Neal (2011) – Repasamos el código de ejemplo de Neal (2011, p.125) línea por línea. La variante `HMC2()` del paquete R `rethinking` proporciona información adicional

como la tasa de aceptación y las trayectorias, de modo que podamos utilizarla posteriormente para demostraciones gráficas del flujo secuencial del HMC. El algoritmo es bastante sencillo. Los parámetros que hay que elegir y las variables a utilizar son (`ptII_quan_Bayes_HMC.r`)

- U = función de la energía potencial, en la práctica una función, que da para q la log-Posterior negativaausgibt
- gradient_U = función de U, que da el gradiente de U en lugar de q (die den Gradienten von U an der Stelle q ausgibt (Derviada parcial de U por q)
- epsilon = el tamaño de paso de los leapfrogs
- L = número de saltos antes de que se proponga un candidato
- current_q = posición actual q
- K = función de la energía cinética con Masa = 1, expresada como $\sum ½p^2$, donde p es la probabilidad de los nuevos valores propuestos.

```
# Radford Neal 2012 R-Code
HMC <- function (U, grad_U, epsilon, L, current_q)
{
```

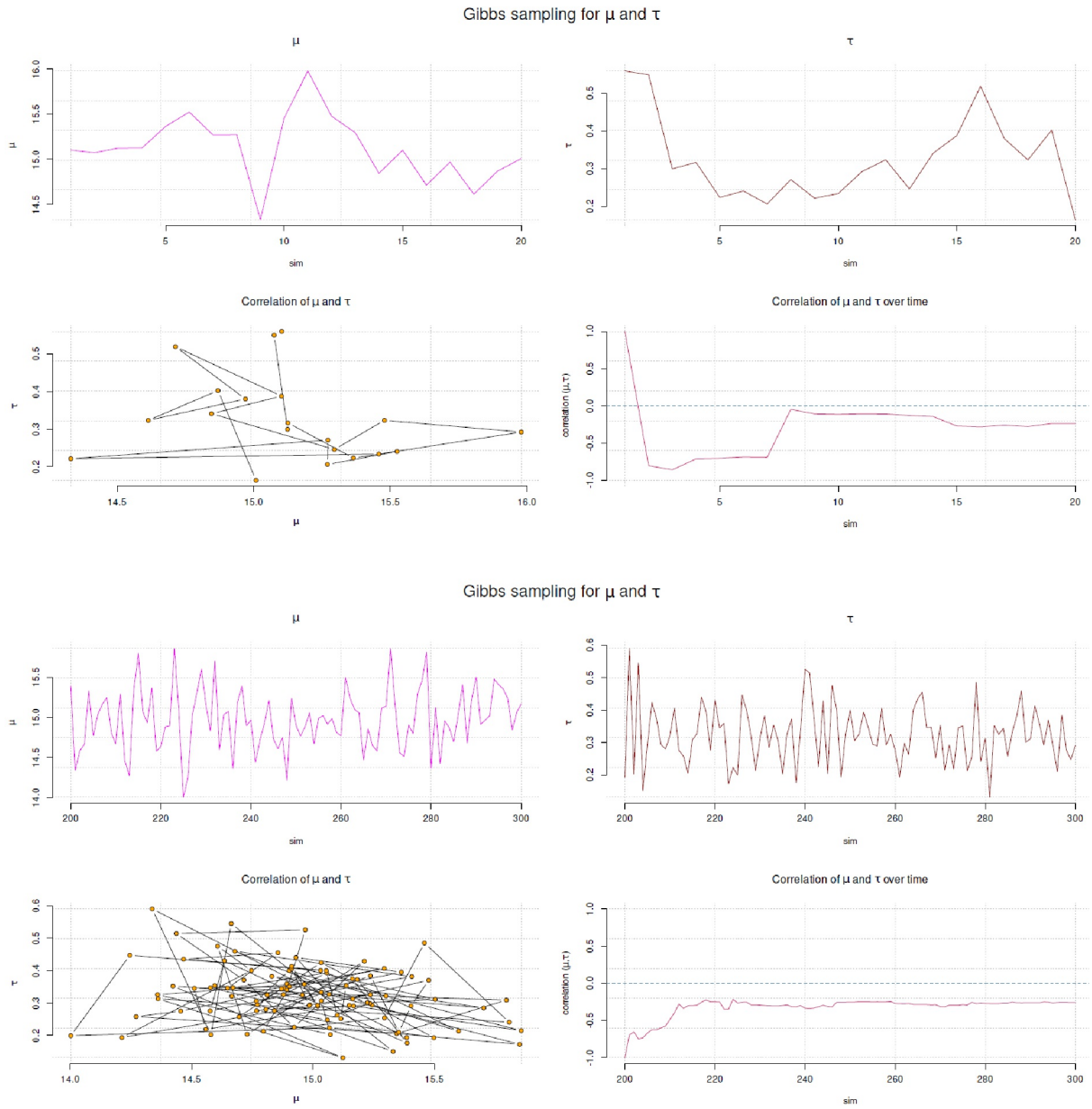


Figura 6.91. *Muestreo de Gibbs (extractos de la secuencia MCMC, simulación 1 a 20 y 200 a 300 respectivamente).*

Comienza con el vector *q* formado a partir del vector actual *actual_q* pasado a la función HMC:

```
q <- current_q
```

Los candidatos como variables aleatorias proceden de una distribución normal estándar independiente con la longitud del vector *q*:

```
p <- rnorm(length(q), 0, 1)
```

En el leapfrog, primero se realiza la separación del paso 1 de Euler y el vector *p* se almacena antes del leapfrog y los cambios asociados como vector *current_p*:

```
current_p <- p
```

Comenzamos con un medio paso para el momento *p*. El medio paso se calcula con el factor ϵ mediante multiplicación y así se obtiene el tamaño real del paso. El cálculo con el factor se aplica a todos los pasos siguientes.

```
p <- p - epsilon * grad_U(q) / 2
```

Alternativamente, un paso entero va para la posición *q* y el momento *p*. El bucle de saltos va de 1 a *L*.

```
para (i en 1:L) {
```

y comienza con un paso entero para la posición *q*.

```
q <- q + epsilon * p
```

Esto es seguido por un paso entero para el momento *p* – salvo al final de la trayectoria de movimiento.

```
If (i != L) p <- p - epsilon * grad_U(q)
}
```

y ahora sigue medio paso al final para el momento *p*

```
p <- p - epsilon * grad_U(q) / 2
```

Para que los candidatos sean simétricos, se invierte el momento (=cambio de signo)

```
p <- -p
```

Con esto concluye el paso leapfrog. El siguiente paso de Metrópolis para la evaluación de los candidatos une *U* con *K* a lo largo de los valores actuales para la posición `current_q` y el momento `current_p` por un lado y los candidatos `proposed_q` y `proposed_p` por otro. Los valores se calculan lógicamente de forma idéntica para cada uno.

```
current_U  <- U(current_q)
current_K  <- sum(current_p^2) / 2
proposed_U <- U(q)
proposed_K <- sum(p^2) / 2
```

La aceptación frente al rechazo de los candidatos tiene lugar de forma clásica según el algoritmo MH frente a un número aleatorio distribuido uniformemente. Se adopta el nuevo vector *q* o bien el vector anterior o se conserva el vector anterior `current_q`. El valor a comprobar se crea en la escala logarítmica formando la suma de las diferencias de `current_U, proposed_U, current_K` y `proposed_K`, que en la escala no logarítmica corresponde al producto de los cocientes respectivos. Este valor debe ser mayor que la variable de comparación aleatoria, el núcleo de la aceptación del algoritmo MH. Si lo es, los candidatos se aceptan como nuevos valores y el y el bucle comienza de nuevo.

El código R de ejemplo de Neal (2011) sólo devuelve el valor de la posición *q*.

```
if (runif(1) < exp(current_U - proposed_U + current_K - proposed_K))
{
    return (q) # accept
    } else {
    return (current_q) # reject
    }
}
```

En la implementación de `HMC2()` en `rethinking`, se emiten las trayectorias de movimiento para *q* y *p*, las tasas de aceptación y la diferencia `dH <- H1-H0`, que pueden definirse de la siguiente manera:

```
H0 <- current_U + current_K R-Code
H1 <- proposed_U + proposed_K
dH <- H1 - H0
```

De este modo, se puede seguir el curso exacto del desarrollo de la distribución simulada y examinar y comparar la eficacia en función de los valores iniciales de *L* y ϵ. También es impresionante ver cómo (véase ejemplo para `HMC_2D_Sampler()` en `rethinking`) se desarrollan los pasos de salto y en qué dirección apuntan (ver Fig.6.94). Con unas pocas simulaciones MCMC se puede obtener una impresión aproximada de cómo una distribución simulada se construye lentamente a través de constantes saltos de leapfrog. En principio, también se podría hacer una pequeña película con esto.

A continuación, utilizamos `HMC2()` del paquete R `rethinking` para demostrar el procedimiento. Para la ilustración se utiliza la simulación de una distribución normal bivariante.

6.13.2.3.2 Estimación de una distribución normal bivariante – Comenzamos con la estimación de una distribución normal bivariante con una matriz Σ de covarianzas conocida y valores medios $\mu_1 = \mu_2 = 0$. En primer lugar, necesitamos la función de densidad de la distribución normal bivariante para $k = 2$ y X como un vector de columna k-dimensional y la correlación ρ entre las dos variables X_1 y X_2, así como las desviaciones estándar $\sigma_1 > 0$ y $\sigma_2 > 0$

$$N(\mu_{k=2}, \Sigma) = \frac{\exp\left(-\frac{1}{2} \cdot (x-\mu)^T \cdot \Sigma^{-1}(x-\mu)\right)}{\sqrt{(2 \cdot \pi)^k \cdot \det(\Sigma)}} \tag{6.142}$$

$$f(x_1, x_2) = \frac{1}{2 \cdot \pi \cdot \sigma_1 \cdot \sigma_2 \cdot \sqrt{1-\rho^2}} \cdot \exp\left(-\frac{1}{2 \cdot (1-\rho^2)} \cdot \left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2 \cdot \rho \cdot \left(\frac{x_1-\mu_1}{\sigma_1}\right) \cdot \left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right) \tag{6.143}$$

con el vector de valores medios μ

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \tag{6.144}$$

y matriz de covarianza Σ

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho \cdot \sigma_1^2 \cdot \sigma_2^2 \\ \rho \cdot \sigma_1^2 \cdot \sigma_2^2 & \sigma_2^2 \end{pmatrix} \tag{6.145}$$

y convertimos la primera fórmula con notación matricial en código R y suponeos que la matriz `sigmamat` es positiva definita e invertible (`ptII_quan_Bayes_HMC.r`):

```
ND.pdf <- function(x, mu, sigmamat)
{
    (2*pi)^(-1) * det(sigmamat)^(-1/2) *
    exp(-1/2 * t(x-mu) %*% solve(sigmamat) %*% (x-mu))
}
```

Ahora vienen los valores teóricos correspondientes con $\mu_1 = \mu_2 = 0$ o $\sigma^{21} = \sigma^{22} = 0$ y $p = 0.8$:

```
# correlation
rho <- 0.8
# sigmas
sigma1 <- 1
sigma2 <- 1
# sigma matrix
sigmamat <- matrix(c(sigma1,rho*sigma1*sigma2,
            rho*sigma1*sigma2,sigma2),ncol=2)
sigmamat

# mu
mu <- c(1,1)
# point where to evaluate function
x <- c(0.5,0.4)
```

A continuación se comprueba si `ND.pdf()` funciona correctamente – comparándolo con `dmvnorm()` del paquete R `mvtnorm` – y se miede el tiempo necesario para este cálculo, ya que el código no está optimizado:

```
> start_time <- Sys.time()
> ND.pdf(x,mu,sigmamat)
[,1]
[1,] 0.2214392
> end_time <- Sys.time()
> end_time - start_time
Time difference of 0.0007381439 secs
> start_time <- Sys.time()
> dmvnorm(x,mean=mu,sigma=sigmamat)
[1] 0.2214392
> end_time <- Sys.time()
> end_time - start_time
Time difference of 0.0008759499 secs
```

Lo mismo ocurre, como debe ser, cuando simplemente se aplican fórmulas. El tiempo necesario también está bien. No hay que sobrestimar las pequeñas diferencias de tiempo, porque el ordenador está constantemente haciendo algo en segundo plano y esto por sí solo puede causar fluctuaciones. Para estar seguro, repite estas mediciones y comprueba si el resultado es más o menos el mismo. Eso es suficiente.

Para implementar el algoritmo HMC, se necesitan los paquetes de R `rethinking` para `HMC2()` y `numDeriv` para calcular el gradiente de la función con `grad()`. Para los diagnósticos MCMC posteriores, se utilizan los paquetes de R `bayesplot` y/o `coda`. Probamos la función gradiente.

Si por casualidad tenemos a mano la derivada de la función básica, por supuesto podemos convertirla directamente en una fórmula, lo que en caso de duda es más rápido en términos de cálculo. La salida de `grad()` se compara con `num_grad()` de `rhmc`:

```
> numDeriv:::grad(f=dmvnorm, x=x, mean=mu, sigma=sigmamat)
[1] 0.01230218 0.12302180
> numDeriv:::grad(f=ND.pdf, x=x ,mu=mu, sigmamat=sigmamat)
```

```
[1] 0.01230218 0.12302180
> rhmc:::num_grad(f=ND.pdf, x=x, mu=mu, sigmamat=sigmamat)
Error en rhmc:::num_grad(f=ND.pdf, x=x, mu=mu, sigmamat=sigmamat):
Argumentos no utilzados (mu=mu, sigmamat=sigmamat)
```

Vemos que `num_grad()` devuelve un error porque la función no nos permite pasar más parámetros a la función gradiente. Podemos arreglar esto fácilmente añadiendo esta capacidad a la función. En R, estos son tres puntos que sirven como marcadores de posición para cualquier parámetro adicional y simplemente lo pasan. Aquí está el fragmento relevante modificado de `num_grad()`:

```
num_grad2 <- function (f, x, ...)
{
    [R-Código omitido]
    g[i] = (f(Xh, ...) - f(x, ...))/dx
    [R-Código omitido]
}
```

Ahora funciona y los valores concuerdan con `grad()`:

```
> num_grad2(f=ND.pdf, x=x, mu=mu, sigmamat=sigmamat)
[1] 0.01230218 0.12302181
```

Ahora podemos pasar a la simulación MCMC. Primero denotamos la función logarítmica negativa de la distribución normal bivariante y su función gradiente. Si no las ha deducido usted mismo, puede hacerlo en R sin ningún problema de acuerdo con el trabajo preparatorio anterior:

```
# we define two functions for the HMC sampling
# negative log likelihood for U
U <- function(q, ...) -mvtnorm:::dmvnorm(x=q, sigma=sigmamat,
       log=TRUE)
# gradient of U by making use of grad() from numDeriv
grad_U <- function(q, ...) -numDeriv:::grad(f=dmvnorm, x=q,
       sigma=sigmamat, log=TRUE)
```

Para que el procedimiento sea repetible, fijamos un valor inicial diferente para el generador de valores aleatorios para cada cadena MCMC, de modo que se pueda reproducir toda la cadena. Hacemos esto para 20 cadenas MCMC y luego sólo utilizamos los valores de salida que son necesarios:

```
seeds <- c(9988776, 996, 345, 321, 12399,
           395, 350, 840, 382, 573,
           242, 891, 385, 680, 606,
           770, 913, 795, 670, 736
           )
```

A continuación se definen los parámetros que controlan el algoritmo HMC. Comenzamos con el número de leapfrogs *L*, los valores de salida de *q* para ambas variables `q1` y `q2`, el número de repeticiones MCMC por cadena MCMC, el número de cadenas MCMC y la covarianza Σ:

```
# epsilon
step <- 0.1
# number of leapfrogs
L <- 11
# initial value for qs at c(0,0)
Qinitv <- c(0,0)
```

```
# number of samples per MCMC chain
nsamp <- 3e4
# number of MCMC chains
nchains <- 5
# covariance sigma
sigmamat
```

Dado que las funciones de R `U()` y `grad_U()` trabajan por defecto con un valor medio de $\mu_1 = \mu_2 = 0$, no es necesario ningún otro valor. Ahora la función R `bivarnormdist.HMC.sim()` lo completa todo, ya que realiza las simulaciones MCMC sobre la base de `HMC2()` de `rethinking`:

```
bivarnormdist.HMC.sim <- function(U, grad_U, R-Code
epsilon=0.1, L=11, Qinitv=c(0,0),
nchains=5, nsamp=1e3,
Qrescnams=c("q1","q2","a","dH"),
       seeds, ...)
{
cat("\nnsamp:\t",nsamp,"\nEpsilon: ",
       epsilon,"\nL:\t",L,"\n\n",sep="")
coln <- length(Qrescnams) #length(Qinit$q)+1+1
Qres <- matrix(NA, nrow=nsamp, ncol=coln)
colnames(Qres) <- Qrescnams
Qres[1,c(1:2)] <- Qinitv
OUTmcmc <- list()
s.l <- length(seeds)
if(s.l != nchains) stop("Number of chains and seeds differ!")
for(z in 1:nchains)
{
cat("Chain:\t",z,"\nseed=\t",seeds[z],"\n\n",sep="")
Q <- list()
Q$q <- Qinitv
set.seed(seeds[z])
for (i in 1:nsamp)
{
# print(i)
Q <- HMC2(U, grad_U, epsilon=step, L=L, current_q=Q$q, ...)
Qres[i,"dH"] <- Q$dH
Qres[i,"a"] <- Q$accept
if(Q$a == 1) Qres[i,c(1:2)] <- Q$q
}
OUTmcmc[[z]] <- Qres
}
return(OUTmcmc)
}
```

De esta manera podemos simular la distribución normal bivariante para diferentes valores iniciales (tamaño del paso ϵ y número deleapfrogs L) y posteriormente hacer comparaciones con otros parámetros de salida. Es importante que la función R utilice un bucle `for()` para procesar el número de repeticiones MCMC. Dado que la tasa de aceptación suele ser inferior a 1, esto hace que la cadena MCMC se acorte automáticamente. En este ejemplo, en combinación con el HMC, esto hace poca diferencia. Más ab ajo en el texto se repite la simulación con el algoritmo MH – y allí hace una gran diferencia si se usa un bucle `for()` sobre el número de repeticiones MCMC o si se denota un bucle con `while()`, que funciona hasta que se haya alcanzado el número deseado de repeticiones MCMC con éxito. En el script R correspondiente `ptII_quan_Bayes_HMC.r` ambas funciones están disponibles, por lo que debería ser fácil reescribir `bivarnormdist.HMC.sim()` con este otro criterio. La llamada a `bivarnormdist.HMC.sim()` es

```
posty3 <- bivarnormdist.HMC.sim(U=U, grad_U=grad_U, epsilon=step, L=11,
           nchains=nchains, Qinitv=Qinitv, nsamp=nsamp,
           seeds=seeds, sigma=sigmamat)
save(posty3, file="posty3_step0.1_L11_nchains5_nsampe34.RDAta")
OUTmcmc <- posty3
```

En primer lugar, almacenamos la salida en la variable `posty3` y la guardamos en disco duro. Esto tiene la ventaja de que no tenemos que repetir la simulación, porque algunas se toman su tiempo. El siguiente paso es pasar la cadena MCMC a una variable uniforme llamada `OUTmcmc`. De esta manera podemos mantener nuestro R-script simple y hacer todos los demás diagnósticos con la variable `OUTmcmc`. Esto se utiliza siempre después de una llamada a `bivarnormdist.HMC.sim()` y la salida se almacena en otra variable (aquí: posty3), para que no se pierda nada. Esta es una buena manera de probar diferentes parámetros. Inspeccionemos la cadena MCMC de `posty3`:

```
> str(posty3)
List of 5
$ : num [1:30000, 1:4] -0.782 -1.023 -1.215 -0.385 0.282 ...
..- attr(*, "dimnames")=List of 2
.. ..$ : NULL
.. ..$ : chr [1:4] "q1" "q2" "a" "dH"
$ : num [1:30000, 1:4] -0.495 0.16 -1.055 0.986 -0.415 ...
..- attr(*, "dimnames")=List of 2
.. ..$ : NULL
.. ..$ : chr [1:4] "q1" "q2" "a" "dH"
$ : num [1:30000, 1:4] -0.5937 0.5772 -0.389 -1.3748 -0.0988 ...
..- attr(*, "dimnames")=List of 2
.. ..$ : NULL
.. ..$ : chr [1:4] "q1" "q2" "a" "dH"
$ : num [1:30000, 1:4] 0.828 -0.385 0.519 -0.444 0.702 ...
..- attr(*, "dimnames")=List of 2
.. ..$ : NULL
.. ..$ : chr [1:4] "q1" "q2" "a" "dH"
$ : num [1:30000, 1:4] -0.345 0.765 0.383 -1.381 -1.142 ...
..- attr(*, "dimnames")=List of 2
.. ..$ : NULL
.. ..$ : chr [1:4] "q1" "q2" "a" "dH"
```

Así pues, tenemos una lista de cinco cadenas MCMC, cada una de las cuales contiene una tabla para las variables `q1, q2, a` y `dH`. El número de filas corresponde al número de repeticiones por cadena MCMC. Los valores *q1* y *q2* son los valores *q* de la distribución apuntada, *a* contiene la tasa de aceptación y *dH* la diferencia `H1-H0` (véanse las explicaciones anteriores del script R de Neal).

Un extracto de la primera cadena tiene el siguiente aspecto:

```
> head(posty3[[1]])
      q1         q2        a  dH
[1,] -0.7822362 -0.4830823 1  0.001707170
[2,] -1.0230008 -0.9611400 1 -0.000887961
[3,] -1.2154996 -0.7183723 1  0.003763706
[4,] -0.3852855 -0.8193120 1 -0.001359583
[5,]  0.2817869  0.8730243 1  0.002495819
[6,]  0.5030277 -0.2462167 1  0.003064933
```

El análisis de la tasa de aceptación es importante. En `HMC2()` está representado por el vector `a` y los valores no aceptados reciben un *NA*. Por lo tanto, se deduce que hay un análisis NA de los valores no aceptados del vector bidimensional `q`. Como sólo hay una prueba de aceptación, `q1` y `q2` reciben cada uno un valor o un *NA*, pero nunca diferentes entre sí.

```
NA.IDs <- lapply(OUTmcmc, function(x) which(is.na(x), arr.ind=TRUE))
NA.rowIDs <- unique(unlist(lapply(NA.IDs, function(x) x[,1])))
NA.rowIDs.l <- length(NA.rowIDs)
```

Observamos la longitud del vector `NA.rowIDs`

```
> NA.rowIDs.l
[1] 82
```

y vemos en la salida de `NA.IDs` que sólo la primera cadena MCMC tiene NAs en absoluto y por lo tanto no aceptaciones. Según `NA.rowIDs.l`, el número es de 82 no aceptaciones. Según 82/30000 = 0,00273

```
> # rate a
> NA.rowIDs.l/nsamp
[1] 0.002733333
```

resulta en un factor de no aceptación inferior al 0,3% en las 30 000 repeticiones. Una tasa de no aceptación del 2,8 % o una tasa de aceptación > 99,7 % es, en efecto, apenas digna de mención... Las demás cadenas MCMC tienen incluso una tasa de aceptación del 100 %. Pero esto no debe tomarse en absoluto como una regla y ciertamente no se aplica a modelos complejos, que no incluyen la simulación de una distribución normal bivariante. No obstante, no podemos utilizar las NA y eliminarlas de la cadena MCMC. Si simulamos hasta un cierto número de repeticiones con éxito, no habría NAs (véase más arriba). Dado que los diagnósticos posteriores suponen que cada cadena MCMC es de igual longitud, reducimos las otras cadenas MCMC en el número de NAs. Esto no es una distorsión en la medida en que las cadenas MCMC son independientes entre sí y, por lo tanto la generación de los NA es aleatoria y no tiene nada que ver con los valores de la cadena MCMC respectiva. Omitimos aquí el código R asociado. También se podrían eliminar aleatoriamente celdas o muestrear hasta que haya el mismo número de valores aceptados en cada cadena.

Si estamos interesados en saber cuántas repeticiones se encuentran típicamente entre las no-aceptaciones y las aceptaciones, esto se puede hacer con

```
> c(summary(NA.rowIDs.diffs), sd=sd(NA.rowIDs.diffs))
Min.   1st Qu.  Median   Mean     3rd Qu.  Max.      sd
1.0000 121.0000 267.0000 369.3086 402.0000 2744.0000 420.6934
```

Así, cada 267 (= mediana) o 369 (= media) repeticiones hay una no-aceptación, con una desviación estándar de 421 repeticiones. Ahora eliminamos las no-aceptaciones y volvemos a comprobar si se ha pasado alguna por alto:

```
OUTmcmc.nonas <- list()
if(length(NA.rowIDs) > 0)
{
for(i in 1:nchains)
{
OUTmcmc.nonas[[i]] <- OUTmcmc[[i]][-NA.rowIDs,c("q1","q2")]
}
} else OUTmcmc.nonas <- OUTmcmc
lapply(OUTmcmc.nonas, function(x) which(is.na(x), arr.ind=TRUE))
```

Este no es el caso, por lo que las cadenas MCMC pueden convertirse a un formato legible por `bayesplot` y `coda` y examinarse gráficamente. Por supuesto, esto sería también posible si los burn-ins se eliminaran de antemano del conjunto de datos:

```
# remove burnins if required
removeburnins <- FALSE
dim(OUTmcmc.nonas[[1]])
burnins <- 500
OUTmcmc.nonas.noburnins <- lapply(OUTmcmc.nonas,
       function(x) x[-c(1:burnins),])
dim(OUTmcmc.nonas.noburnins[[1]])
if(removeburnins)
{
```

```
OUTmcmc <- OUTmcmc.nonas.noburnins else OUTmcmc <- OUTmcmc.nonas
}
```

No lo hacemos explícitamente porque nos interesa el principio de la cadena MCMC y no el curso estacionario posterior, donde hay poco que aprender. Por lo demás, el análisis se concentra en `q1` y `q2`, de modo que `a` y `dH` se eliminan de antemano.

```
# everything R-Code
OUTmcmc.nonas.onlyqs <- OUTmcmc.nonas
# only q's
OUTmcmc.nonas.onlyqs <- lapply(OUTmcmc.nonas,
                         function(x) x[,c("q1","q2")])
```

A esto le sigue la preparación para `coda`

```
OUTmcmc.list <- as.mcmc.list(lapply(OUTmcmc.nonas.onlyqs, mcmc))
```

y un resumen de las cadenas MCMC:

```
> summary(OUTmcmc.list)
Iterations = 1:29918
Thinning interval = 1
Number of chains = 5
Sample size per chain = 29918
1. Empirical mean and standard deviation for each variable,
        plus standard error of the mean:
   Mean    SD    Naive SE  Time-series SE
q1 0.00333 0.999 0.00258   0.00565
q2 0.00347 0.999 0.00258   0.00564
2. Quantiles for each variable:
    2.5%  25%  50%      75%   97.5%
q1 -1.96 -0.67 0.004145 0.679 1.96
q2 -1.96 -0.67 0.000687 0.680 1.96
```

Las estadísticas descriptivas son interesantes para las cadenas MCMC individuales, con el fin de ver las diferencias en cada caso. Para ello utilizamos la función de R adaptada a nuestro caso `MCMCout.desc.per.chain()`, que funciona por cadena MCMC

```
> MCMCout.desc.per.chain(OUTmcmc.nonas.onlyqs, nchoose=c(1,2))
$q1
      Min.   1st Qu. Median   Mean    3rd Qu. Max.  sd   var
[1,] -3.72   -0.648  0.01659  0.02258 0.701   4.21 0.999 0.998
[2,] -4.13   -0.670  0.01408  0.00951 0.686   3.71 1.005 1.009
[3,] -4.22   -0.663  0.01156  0.00172 0.678   3.80 0.996 0.993
[4,] -4.22   -0.685 -0.00779 -0.00794 0.668   3.87 1.001 1.002
[5,] -4.81   -0.686 -0.01467 -0.00922 0.665   4.19 0.994 0.989
$q2
      Min.   1st Qu. Median   Mean    3rd Qu. Max.  sd   var
[1,] -4.02   -0.651  0.02236  0.02299 0.701   3.89 0.997 0.994
[2,] -4.51   -0.665  0.00464  0.01048 0.687   3.86 1.006 1.013
[3,] -4.44   -0.670  0.00248  0.00239 0.682   3.82 0.998 0.996
[4,] -4.26   -0.683 -0.01158 -0.00969 0.669   4.59 1.003 1.005
[5,] -4.22   -0.685 -0.01601 -0.00884 0.658   4.97 0.992 0.985
```

y `MCMCout.desc.all.chain()`, que recorre todas las cadenas MCMC:

```
> MCMCout.desc.all.chain(OUTmcmc.nonas.onlyqs)
    Min. 1st Qu. Median   Mean    3rd Qu. Max. sd    var
q1 -4.81 -0.67   0.004145 0.00333 0.679   4.21 0.999 0.998
q2 -4.51 -0.67   0.000687 0.00347 0.680   4.97 0.999 0.999
```

Los cuantiles complementan la vista, de nuevo por cadena MCMC

```
> # quantiles per chain
> quans <- c(0,0.05,0.1,0.25,0.5,0.75,0.87,0.9,0.95,0.99,1)
> do.call("rbind", lapply(OUTmcmc.nonas.onlyqs,
+ quantile, probs=quans))
      0%    5%    10%   25%    50%     75%   87%  90%  95%  99%  100%
[1,] -4.02 -1.62 -1.26 -0.649  0.01948 0.701 1.15 1.30 1.66 2.35 4.21
[2,] -4.51 -1.65 -1.28 -0.667  0.00945 0.686 1.14 1.30 1.66 2.33 3.86
[3,] -4.44 -1.65 -1.29 -0.666  0.00647 0.680 1.12 1.27 1.63 2.29 3.82
[4,] -4.26 -1.66 -1.29 -0.684 -0.01031 0.668 1.12 1.27 1.63 2.33 4.59
[5,] -4.81 -1.63 -1.28 -0.685 -0.01548 0.661 1.12 1.27 1.63 2.29 4.97
```

y a través de todas las cadenas MCMC:

```
> # quantiles over all chains
> apply(do.call("rbind", OUTmcmc.nonas.onlyqs),2,quantile, probs=quans)
      q1       q2
0%   -4.81238 -4.510109
5%   -1.64574 -1.638711
10%  -1.28338 -1.276597
25%  -0.67031 -0.669941
50%   0.00415  0.000687
75%   0.67945  0.679874
87%   1.12799  1.130413
90%   1.28165  1.283460
95%   1.64124  1.644002
99%   2.32544  2.319955
100%  4.20608  4.965572
```

Ahora queda analizar la matriz de covarianza, por cadena MCMC

```
> lapply(OUTmcmc.nonas.onlyqs, cov)
[[1]]
    q1      q2
q1 0.998 0.799
q2 0.799 0.994
[[2]]
    q1      q2
q1 1.009 0.806
q2 0.806 1.013
[[3]]
    q1      q2
q1 0.993 0.796
q2 0.796 0.996
[[4]]
    q1      q2
q1 1.002 0.801
q2 0.801 1.005
[[5]]
    q1      q2
q1 0.989 0.790
q2 0.790 0.985
```

y a través de todas las cadenas MCMC:

```
> # covariance matrices over all chains
> cov(do.call("rbind", OUTmcmc.nonas.onlyqs))
    q1     q2
q1 0.998 0.799
q2 0.799 0.999
```

De los gráficos (Fig. 6.92) proporcionados por el paquete `bayesplot` de R, el gráfico de traza con `mcmc_trace()` y la estimación de densidad por cadena MCMC con `mcmc_dens_overlay()` valen la pena, ya que proporcionan una buena imagen general tanto del curso como de la ubicación de la distribución simulada resultante. Los gráficos de violín con `mcmc_violin()`, por otro lado, resultan ser una alternativa refrescante a los boxplots:

```
# trace plot
color_scheme_set("mix-blue-pink")
mcmc_trace(OUTmcmc.nonas)
# density lines for each chain
color_scheme_set("blue")
mcmc_dens_overlay(OUTmcmc.nonas)
# violine plot
color_scheme_set("teal")
mcmc_violin(OUTmcmc.nonas, probs=c(0.1, 0.5, 0.9))
```

Como conclusión preliminar, antes de entrar en más detalles sobre los cambios de los parámetros, está el gráfico bivariante de la distribución normal basada en los valores de simulación generados (véase la Fig. 6.93). Se utiliza una única cadena MCMC y se dibujan elipses con $\alpha = 0{,}05$ (confianza) para los valores esperados teóricamente y los generados empíricamente utilizando `ellipse()` del paquete `mixtools` de R. A esto se añaden $m = 1000$ valores aleatorios simulados de la distribución normal multivariante con `rmvnorm()` del paquete R `mvtnorm` (aquí para dos dimensiones) más la elipse correspondiente con la misma confianza.

```
# plot bivariate normal dist plot with ellipses R-Code
# take on MCMC chain
post <- OUTmcmc.nonas.onlyqs[[1]]
method <- "dmvnorm"
# calculate limits for the plot
limits <- ceiling(abs(apply(post,2,range)))
limits[1,] <- limits[1,]*(-1)
colnams <- colnames(post)
# theoretical mean
mu <- c(0,0)
# empirical mean and covariance matrix
Xbar <- apply(post,2,mean)
S <- cov(post)
plot(NULL, xlim=limits[,"q1"], ylim=limits[,"q2"],
       pre.plot=grid(), bty="n",
       xlab=colnams[1], ylab=colnams[2])
points(post, col="olivedrab", bg="orange", pch=21, cex=0.7)
# draw ellipse based on empirical values
mixtools:::ellipse(mu=Xbar, sigma=S, alpha=0.05,
       col="blue", lwd=2)
# draw ellipse based on theoretical values
mixtools:::ellipse(mu=mu, sigma=sigmamat, alpha=0.05,
       col="magenta", lwd=2)
```

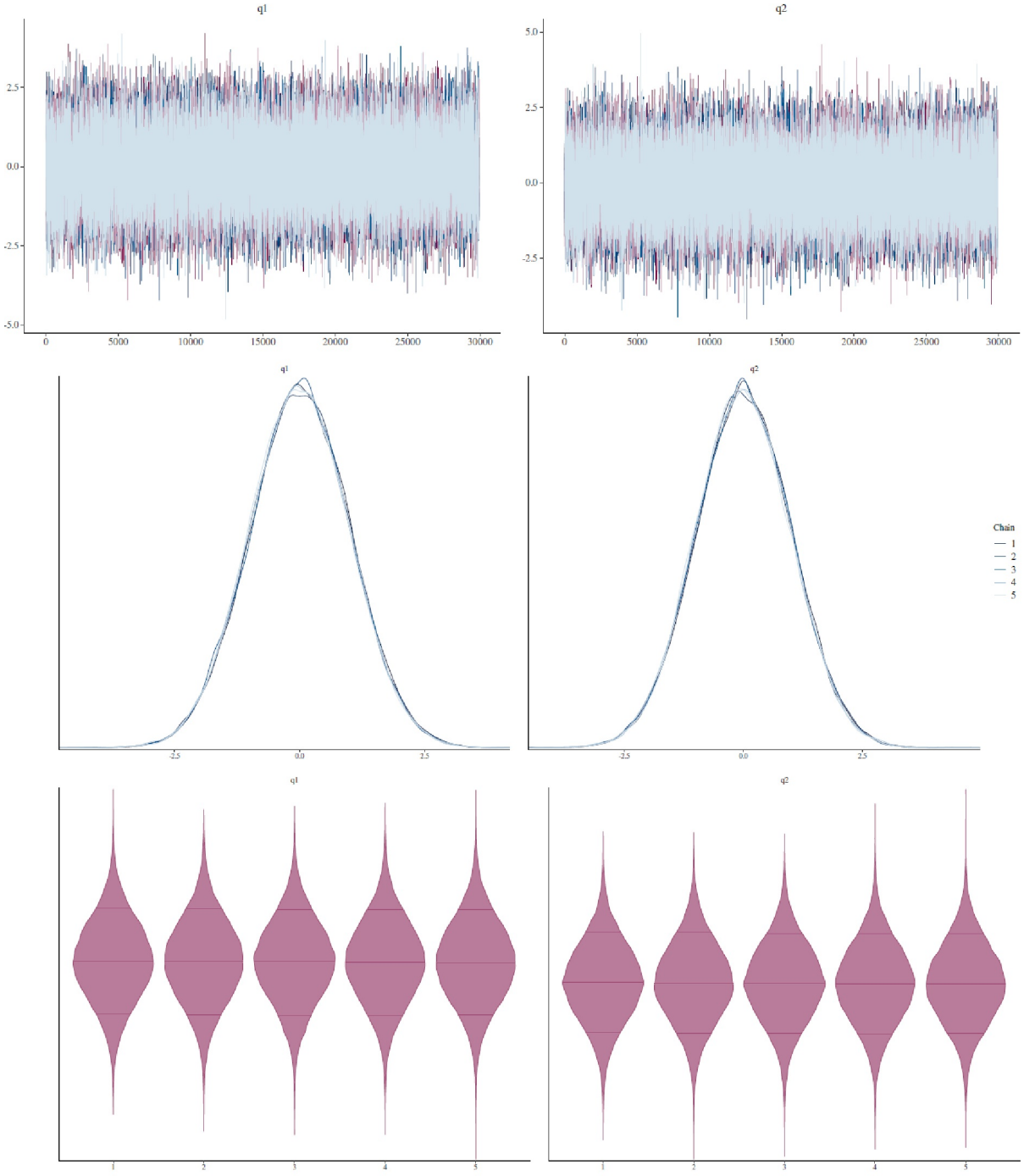


Figura 6.92. *Simulación HMC distribución normal bivariante (trazado, densidades, trazado de violín MCMC)*

Como contraste, los valores de una distribución normal multivariante simulada con `rmvnorm()` del paquete R `mvtnorm`:

```
# compare with simulated from multivariate normal distribution
nsim <- 1e3
rmv.sim <- mvtnorm:::rmvnorm(n=nsim, mean=mu, sigma=sigmamat)
# empirical mean and covariance matrix
Xbar.sim <- apply(rmv.sim, 2, mean)
```

```
S.sim <- cov(rmv.sim)
points(rmv.sim, col="skyblue", bg="yellow", pch=21, cex=0.7)
# draw ellipse based on empirical values
mixtools:::ellipse(mu=Xbar.sim, sigma=S.sim, alpha=0.05,
        col="black", lwd=2, lty=2)
mtext("Bivariate normal distribution (HMC simulation)",
        side=3, cex=2)
```

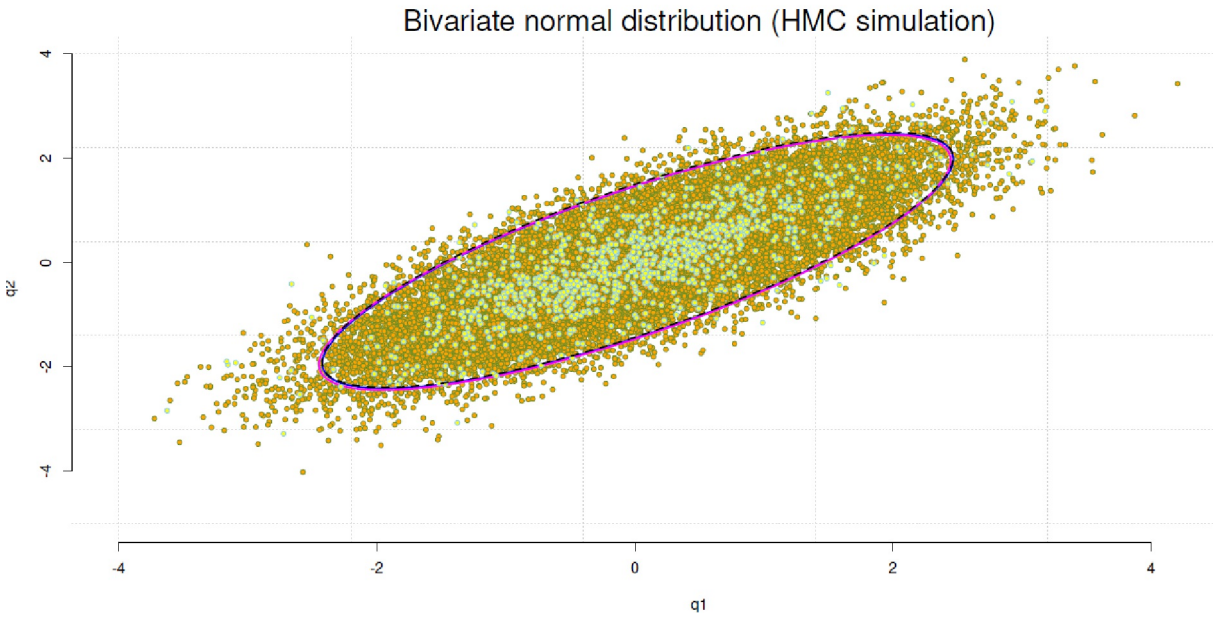


Figura 6.93. *Distribución normal bivariante de simulación HMC (distribución simulada con elipses)*

Como se puede ver en la Figura 6.93, las tres elipses se solapan más o menos completamente y los valores aleatorios de la de la simulación con `rmvnorm()` encajan perfectamente en las simulaciones HMC. Por lo tanto, la simulación HMC ha aproximado la distribución apuntada con mucha precisión. Las estadísticas descriptivas anteriores para `q1` y `q2` así como la covarianza de `q1` y `q2` no sugieren nada más.

La construcción inicial de la cadena MCMC puede observarse con `HMC_2D_Sample()` del paquete R `rethinking`. La Figura 6.94 lo muestra para n = {5; 15; 25; 99; 100} muestras. Las muestras no son reproducibles de esta manera, ya que tendríamos que cambiar el código fuente de `HMC_2D_Sample()` para permitir el paso reproducible del valor de salida del generador aleatorio. A partir de un tamaño de muestra de n = 100, el gráfico ya no muestra las trayectorias, sino sólo los valores finales de `q`. Para que la salida gráfica no se ejecute simplemente sin pausa, se utiliza `par(ask=TRUE)` para esperar una entrada cada vez antes de que aparezca el siguiente gráfico.

```
par(ask=TRUE)
samplesn <- c(5,15,25,99,100)
for(i in samplesn)
{
HMC_2D_sample(n=i, U=U, U_gradient=grad_U,
        step=step, L=L, start=Qinitv,
        sigma=sigmamat)
}
```

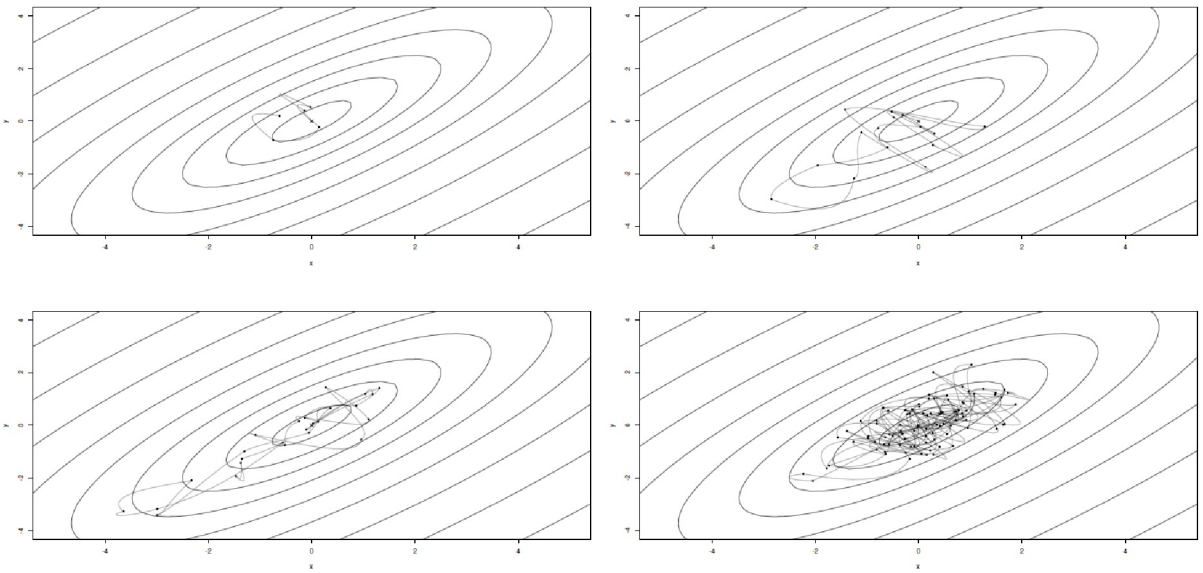


Figura 6.94. *Simulación HMC distribución normal bivariante*
(desarrollo de MCMC para n = {5; 15; 25; 99})

Lo que ahora puede parecer importante es la cuestión del desarrollo de la propia cadena MCMC, es decir, lo rápido o lento y en qué desvíos cambian el valor medio y la matriz de covarianza. El objetivo es, por supuesto, la convergencia de la cadena MCMC. Para ello, las no-aceptaciones se desvanecen, ya que no proporcionan ningún valor perteneciente a la distribución apuntada. Sin embargo, sería concebible almacenar también los valores no aceptados y hacerlos visibles en el gráfico.

Como puede verse a continuación, dependiente de los parámetros iniciales pueden transcurrir distintos periodos de tiempo hasta que se produce la convergencia y para modelos complejos este proceso de convergencia también puede fallar. Para visualizar el proceso de evolución de una cadena MCMC, se requiere una función R que determina y almacena el valor medio y la matriz de covarianza para cada repetición de la cadena MCMC. Como sólo nos interesan las covarianzas, basta con almacenar un único valor. Las varianzas no son de interés aquí. Los valores medios se almacenan para `q1` y `q2` respectivamente. La función de R `MCMCout.cumdesc.per.chain()` se encarga de esta tarea:

```
MCMCout.cs.descs <- lapply(OUTmcmc.nonas.onlyqs,
                           MCMCout.cumdesc.per.chain)
str(MCMCout.cs.descs)
```

La cuestión que se plantea es que a veces sólo se quiere ver una parte de la cadena MCMC. Seguimos necesitando una función de R que ponga a nuestra disposición sólo partes de la cadena y por elemento de la lista, es decir, por cadena MCMC. Esto se puede hacer con una simple llamada a `lapply()` y la selección de elementos del vector con `vector[start:end]`. En el siguiente caso, el inicio y el final del vector de interés se definen primero de 500 (= burn-in) a 2500. Este valor se pasa a `lapply()`. El primer elemento de la lista contiene los valores de `q1` y `q2`. Entonces sólo se selecciona `q1`, ya que los gráficos se dibujan individualmente para cada variable.

```
outtake.q1q2 <- c(start=500,end=2500)
daten <- unlist(lapply(MCMCout.cs.descs,
        function(x) x[[1]][,"q1"][outtake["start"]:outtake["end"]])
)
```

Dado que puede ocurrir que los estadísticos de resumen de la cadena MCMC estén bastante alejados del valor de referencia, es decir, de los valores teóricos para los valores medios y la covarianza, los límites del gráfico deben ajustarse antes de la salida gráfica de modo que el valor de referencia aparezca siempre visiblemente en el gráfico. La función de R `adj.limits()` espera los datos y un valor de referencia `comp.v` y ajusta el rango de los datos en consecuencia. Estos límites especificados se pasan a la variable `ylim` en la llamada a `plot()`:

```
lengs <- sapply(MCMCout.cs.descs[[1]],length)
# length for q1, q2
q1q2.length <- lengs[1]/length(mu)
# length for cov matrix
cov.length <- lengs[2]
q1q2.length
cov.length
# q1
outtake <- outtake.q1q2
comp.value <- mu[1]
listelement <- 1
dura <- outtake["start"]:outtake["end"]
daten <- unlist(lapply(MCMCout.cs.descs,
        function(x) x[[1]][,"q1"][dura])
)
range(daten, na.rm=TRUE)
ylimits <- adj.limits(daten, comp.v=comp.value)
plot(NULL, xlim=c(outtake["start"],outtake["end"]),
        ylim=ylimits,
        type="l", bty="n", pre.plot=grid(col="gray"),
        ylab="q1", xlab="nsamp")
for(i in 1:nchains)
{
lines(outtake["start"]:outtake["end"],
        MCMCout.cs.descs[[i]][[listelement]][,"q1"][dura],
        col=colos1[i])
}
abline(h=comp.value, col="black", lty=2)
mean.per.nsamp <- apply(sapply(MCMCout.cs.descs,
        function(x) x[[1]][,"q1"]),1,mean)
lines(outtake["start"]:outtake["end"],
        mean.per.nsamp[dura], col="red", lwd=3)
```

Se sigue un procedimiento equivalente para q2 y la covarianza. La figura 6.95 muestra los gráficos para `q1, q2` y las covarianzas para el intervalo sin burn-ins de $n_i = 500$ a $n_i = 2500$. La figura 6.96 muestra el final de esta distribución desde $n_i = 25000$ hasta el final de la cadena MCMC. No hay que dejarse engañar por las fluctuaciones "obvias", sino tomar el eje Y y examinar de cerca el tamaño de la fluctuación. Las cadenas MCMC individuales se alejan del valor de referencia 0 un máximo de 0,03 mientras que toda la cadena MCMC (= línea roja) está prácticamente en los valores de referencia para `q1,q2` y la covarianza y fluctúa muy poco a su alrededor. La cadena MCMC completa puede verse en la figura (6.97). Otro vistazo a un gráfico de trazas para la variable `dH` (véase la figura 6.98) no muestra anomalías:

```
# investigate dH R-Code
dH.list <- lapply(OUTmcmc, function(x) x[,"dH"])
str(dH.list)
colos.r <- rainbow(nchains)
plot(dH.list[[i]], type="l", bty="n",pre.plot=grid(),
        col=colos.r[1], xlab="nsamp", ylab="dH")
for(i in 2:nchains) points(dH.list[[i]], type="l", col=colos.r[i],)
```

Esto se aplica igualmente a las estadísticas descriptivas:

```
> do.call("rbind", lapply(dH.list,
+ function(x) c(summary(x),sd=sd(x),var=var(x))))
      Min.    1st Qu. Median   Mean     3rd Qu. Max.   sd      var
[1,] -0.0586 -0.00299 7.75e-06 3.20e-05 0.00301 0.0955 0.00797 6.35e-05
[2,] -0.0694 -0.00307 1.14e-06 2.79e-05 0.00307 0.0741 0.00804 6.46e-05
[3,] -0.0572 -0.00301 1.11e-05 3.41e-05 0.00308 0.0673 0.00791 6.26e-05
[4,] -0.0749 -0.00300 3.68e-07 3.23e-05 0.00305 0.0703 0.00803 6.44e-05
[5,] -0.0613 -0.00296 3.71e-05 3.25e-05 0.00306 0.0567 0.00787 6.19e-05
```

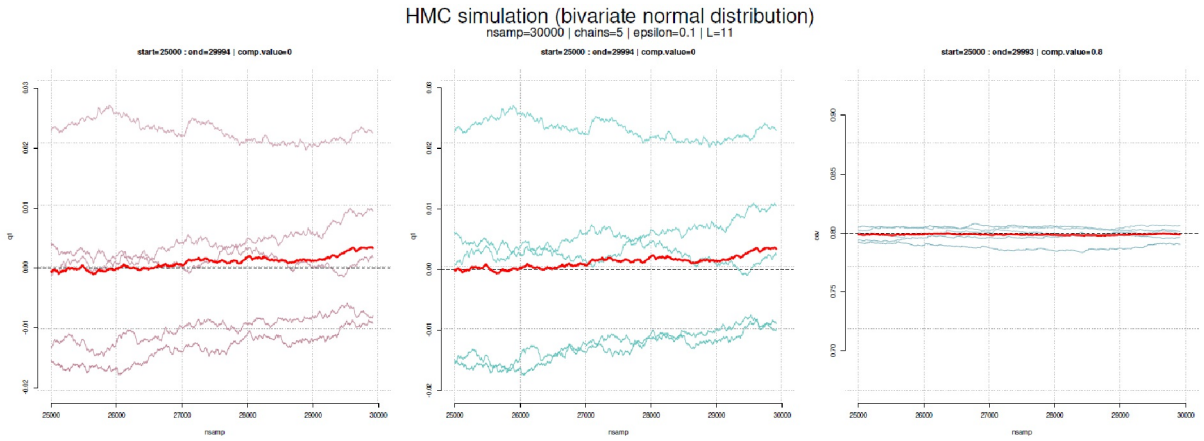


Figura 6.96. *Simulación HMC distribución normal bivariante*
(evolución del MCMC[25000:end] para q1, q2 y covarianza, ϵ = 0,1, L = 11).

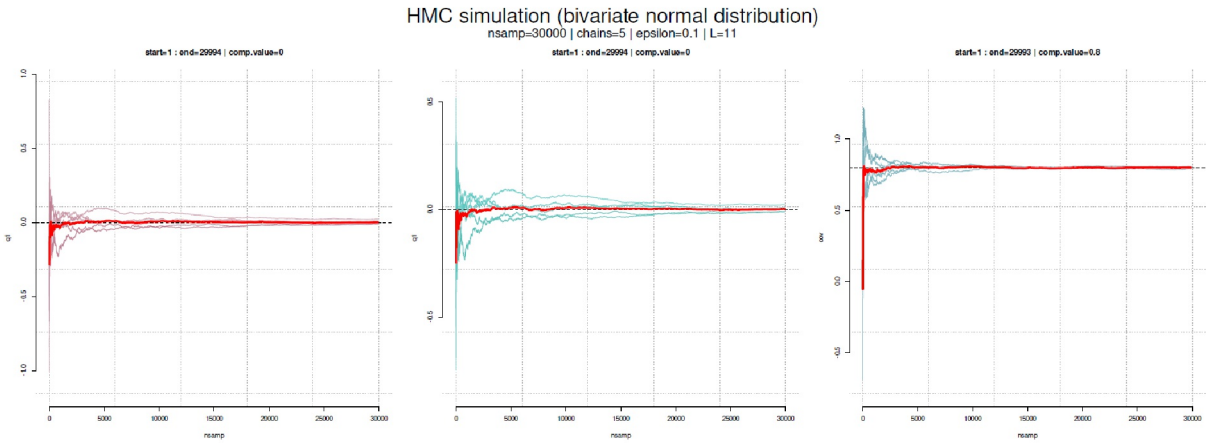


Figura 6.97. *Distribución normal bivariante de simulación HMC*
(evolución de MCMC [1:end] para q1, q2 y covarianza, ϵ = 0,1, L = 11).

Los valores iniciales del algoritmo HMC son importantes. Por ejemplo, el mismo ejemplo de la distribución normal bivariante muestra que con diferentes parámetros se tarda bastante más en que las cadenas MCMC converjan. La Figura 6.99 contiene los gráficos para `q1, q2` y la covarianza para donde sólo se eligió $\epsilon = 0{,}03$ en lugar de $\epsilon = 0{,}1$. Todos los demás parámetros, incluidos los valores de salida del generador aleatorio se mantienen constantes. Para ilustrarlo se utilizan $m = 20$ cadenas MCMC, cada una de las cuales contiene sólo $k = 1000$ repeticiones. La dirección es suficiente para la demostración. Los valores perdidos, es decir, los no-aceptados, no están presentes.

El resumen con el paquete `coda` de R tiene buena pinta:

```
> summary(OUTmcmc.list)
Iterations = 1:1000
Thinning interval = 1
Number of chains = 20
Sample size per chain = 1000

1. Empirical mean and standard deviation for each variable,
    plus standard error of the mean:
     Mean    SD    Naive SE Time-series SE
q1 -0.0648 0.977 0.00691  0.0447
q2 -0.0536 0.990 0.00700  0.0461


2. Quantiles for each variable:
     2.5%  25%    50%    75%   97.5%
q1 -1.95 -0.742 -0.0800 0.606 1.86
q2 -2.02 -0.716 -0.0494 0.604 1.91
```

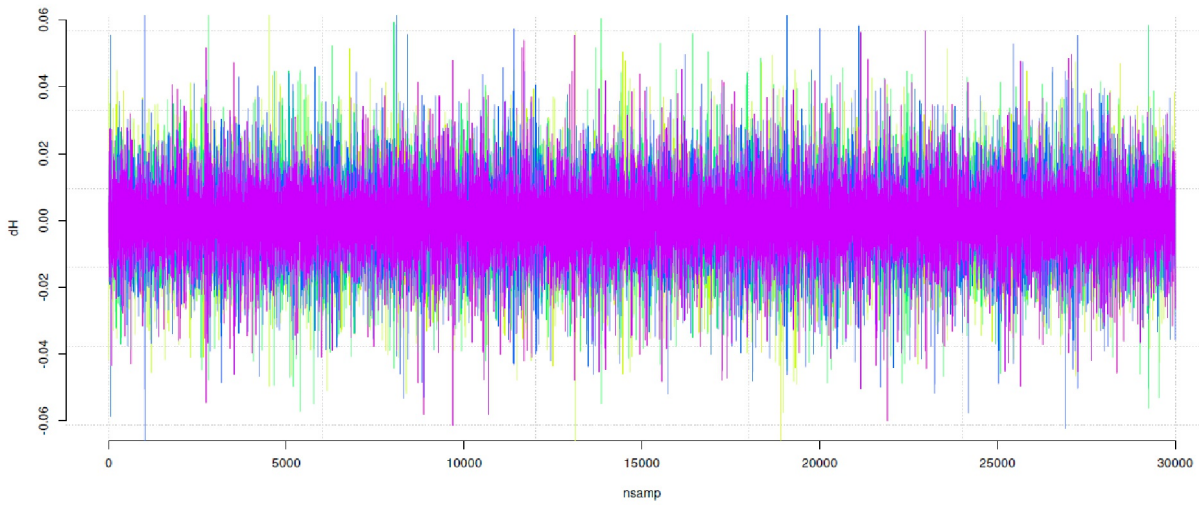


Figura 6.98. *Distribución normal bivariante de la simulación HMC (traceplot dH, ϵ = 0,1, L = 11)*

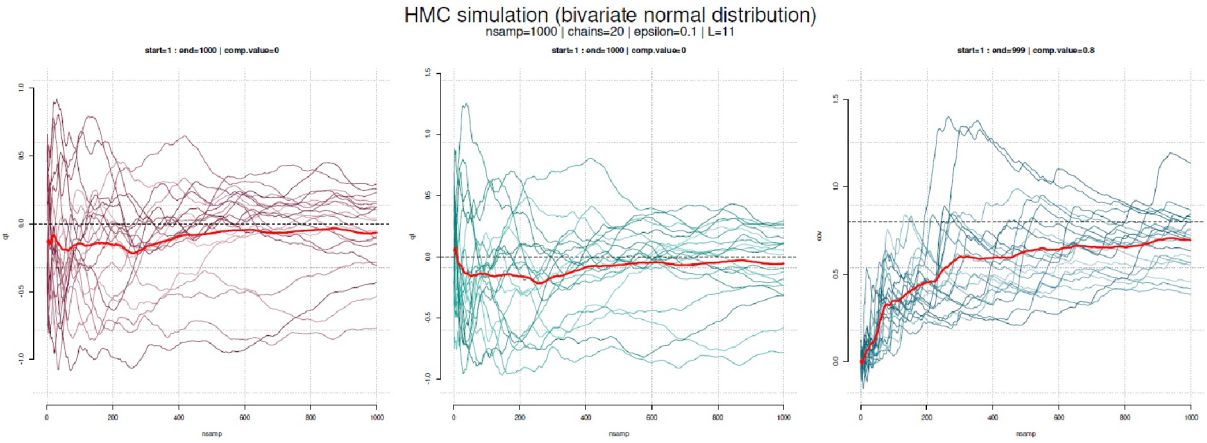


Figura 6.99. *Simulación HMC distribución normal bivariante*
(evolución de MCMC[1:end] para q1, q2 y covarianza, ϵ = 0,03, L = 11)

Pero si ahora miramos las cadenas MCMC individuales, aquí los valores medios, las cosas parecen diferentes:

```
> lapply(MCMCout.desc.per.chain(OUTmcmc.nonas.onlyqs,
        + nchoose=c(1,2)), function(x) x[,"Mean"])
$q1
[1]  0.0442 -0.0519  0.0300 -0.0996 -0.5342  0.1700 -0.1099 -0.7676
[9] -0.2800  0.1039  0.2293  0.1527  0.2945 -0.0934 -0.1475 -0.3052
[17] 0.1206  0.1226 -0.4325  0.2568
$q2
[1]  0.1079  0.0567  0.0241 -0.1214 -0.5743  0.2407 -0.2290 -0.7781
[9] -0.2304  0.1114  0.2969  0.1161  0.2706 -0.1796 -0.0920 -0.3102
[17] 0.1537  0.1167 -0.3134  0.2620
```

El rango se aleja bastante en ambas direcciones de 0, el valor de referencia:

```
> lapply(mcmc.perchain.means, range)
$q1
[1] -0.768 0.295
$q2
[1] -0.778 0.297
```

Mientras que la totalidad de las cadenas MCMC ya se aproxima bien al valor de referencia en n = 1000 repeticiones sin eliminar los burn-ins, las cadenas MCMC individuales muestran lo contrario. Lo mismo ocurre con la covarianza (salida no impresa). Expresado gráficamente, esto se nota en la Figura 6.100 como un gráfico de densidad, que muestra la heterogeneidad en este punto de tiempo.

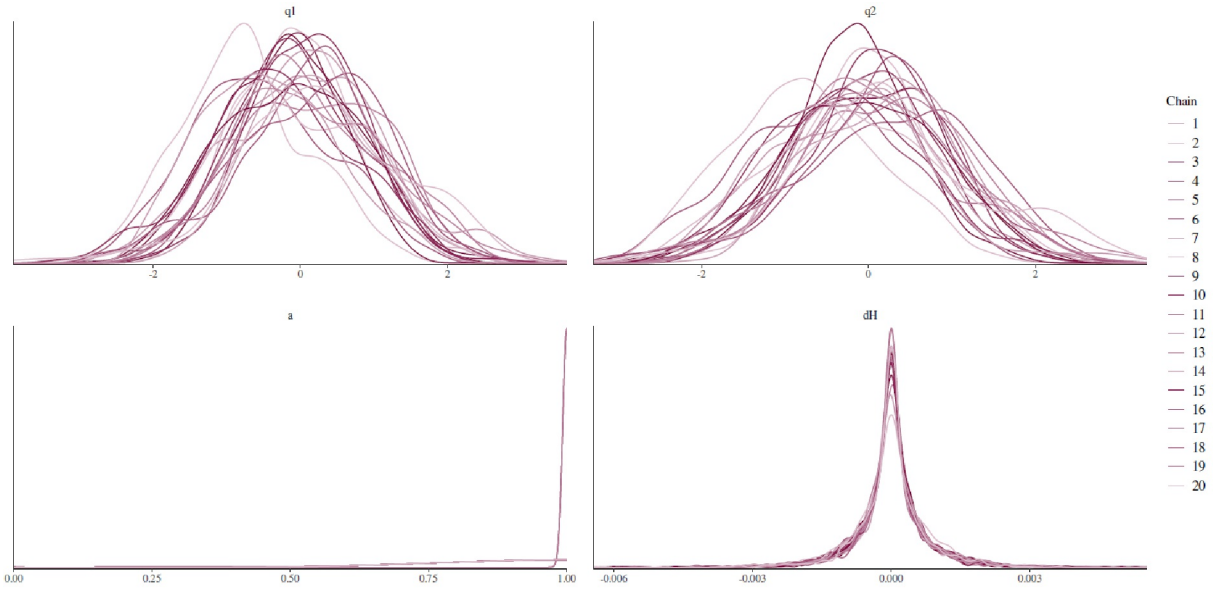


Figura 6.100. *Distribución normal bivariante de simulación HMC*
(densidad estimada `q1, q1, a` *y* `dH`*,* $\epsilon = 0{,}03$*,* $L = 11$*)*

Con el valor $\epsilon = 0{,}1$ y dado las mismas condiciones generales vemos una imagen diferente:

```
> summary(OUTmcmc.list)
Iterations = 1:994
Thinning interval = 1
```

```
Number of chains = 20
Sample size per chain = 994
1. Empirical mean and standard deviation for each variable,
        plus standard error of the mean:
    Mean    SD    Naive SE Time-series SE
q1 -0.0161 0.986 0.00699  0.0149
q2 -0.0146 0.990 0.00702  0.0151
2. Quantiles for each variable:
    2.5%  25%    50%    75%   97.5%
q1 -1.94 -0.684 -0.0202 0.653 1.91
q2 -1.94 -0.690 -0.0154 0.659 1.93
```

y además

```
> mcmc.perchain.means <- lapply(
        + MCMCout.desc.per.chain(OUTmcmc.nonas.onlyqs,
        + nchoose=c(1,2)), function(x) x[,"Mean"])
> mcmc.perchain.means
$q1
[1]   0.03260 -0.00821 0.01409 -0.02120 -0.15312 0.06274 -0.04653
[8]  -0.22650 -0.07693 0.02855  0.07228  0.04218 0.08828 -0.03313
[15] -0.03547 -0.10105 0.03636  0.03651 -0.11067 0.07755
$q2
[1]   0.04066  0.00632 0.01332 -0.02627 -0.15700 0.07014 -0.06076
[]   -0.22758 -0.06997 0.02720  0.08136  0.03403 0.08540 -0.04225
[15] -0.02930 -0.10085 0.03682  0.03978 -0.09333 0.07930
> lapply(mcmc.perchain.means, range)
$q1
[1] -0.2265 0.0883
$q2
[1] -0.2276 0.0854
```

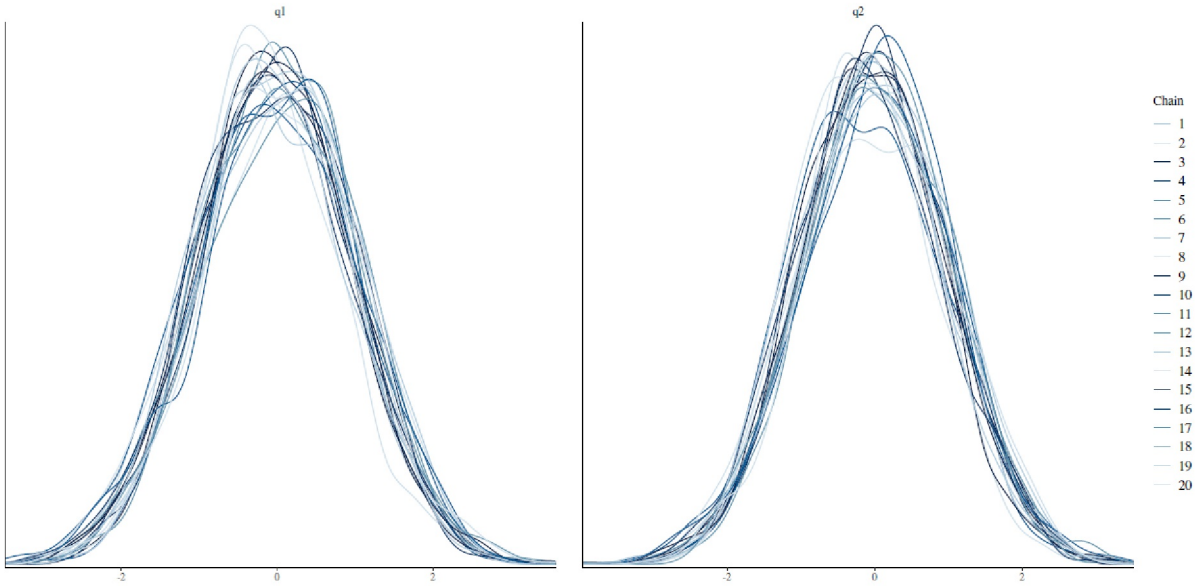


Figura 6.101. *Distribución normal bivariante de la simulación HMC*
(diagrama de densidad `q1,q1,a` *y* `dH,`$\epsilon = 0{,}03$, $L = 11$*)*

El gráfico de densidad (véase la Figura 6.101) ya muestra un solapamiento mucho mayor que el de la Figura 6.100. La Figura 6.102 ilustra la rapidez con que la totalidad de las cadenas MCMC converge hacia

los valores teóricos previstos. Esto contrasta completamente con la Figura 6.99, donde se tarda mucho más y las cadenas MCMC individuales muestran desviaciones masivas del valor apuntado durante un periodo de tiempo más largo (= número de repeticiones).

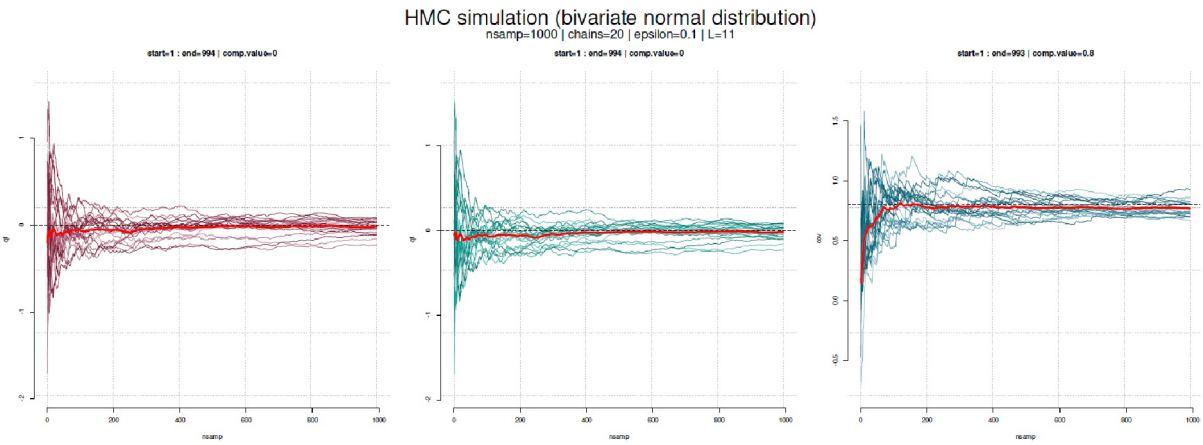


Figura 6.102. *Simulación HMC distribución normal bivariante*
(evolución de MCMC [1:end] para `q1, q2` *y covarianza,* $\epsilon = 0{,}1$, $L = 11$*)*

Como se puede ver, en general, sigue habiendo una aproximación razonable en todas las cadenas MCMC y en todas las repeticiones. Pero con valores iniciales pobres se tarda bastante más. Sin embargo, simular una distribución normal bivariante no es una cosa demasiado compleja. Sólo hay dos dimensiones y la ecuación de densidad básica no es especialmente complicada, en comparación con modelos estadísticos como los HLM, que tienen que estimar muchos parámetros y, por tanto, son de alta dimensión. Entonces, la elección de parámetros HMC desfavorables junto con una Prior igualmente desfavorable marca una gran diferencia y no es sorprendente que un modelo de este tipo falle. Entonces ya no se trata *sólo* de una cuestión de duración.

La figura 6.103 (arriba para [500:2500], centro para [25000:end] y abajo para [1:end]) resume para el caso con $\epsilon = 0{,}03$, 30 000 repeticiones y 5 cadenas MCMC. [1:end]) la evolución para los parámetros `q1, q2` y covarianza.

Los valores iniciales desfavorables del generador de valores aleatorios pueden dar lugar inicialmente a progresiones desfavorables. Sin embargo, éstas se equilibran con el tiempo y, en el peor de los casos, tardan un poco más. Esto se muestra en las figuras 6.102 y 6.99. La razón es que, debido a los diferentes generadores de valores aleatorios, el curso a través del espacio de parámetros es diferente y el algoritmo aparentemente "se pierde" durante bastante tiempo antes de converger a los valores esperados.

Esta es la razón por la que existe una fase de burn-in. El algoritmo HMC elimina todos estos problemas iniciales con el tiempo. Lo que no puede hacer, en cambio, es eliminar los valores iniciales mal ajustados que prácticamente impiden la convergencia. Así que tengamos en cuenta lo que es importante en HMC y lo que puede ir mal – la respuesta general es por un lado valores optimizados y por otro lado valores iniciales desfavorables.

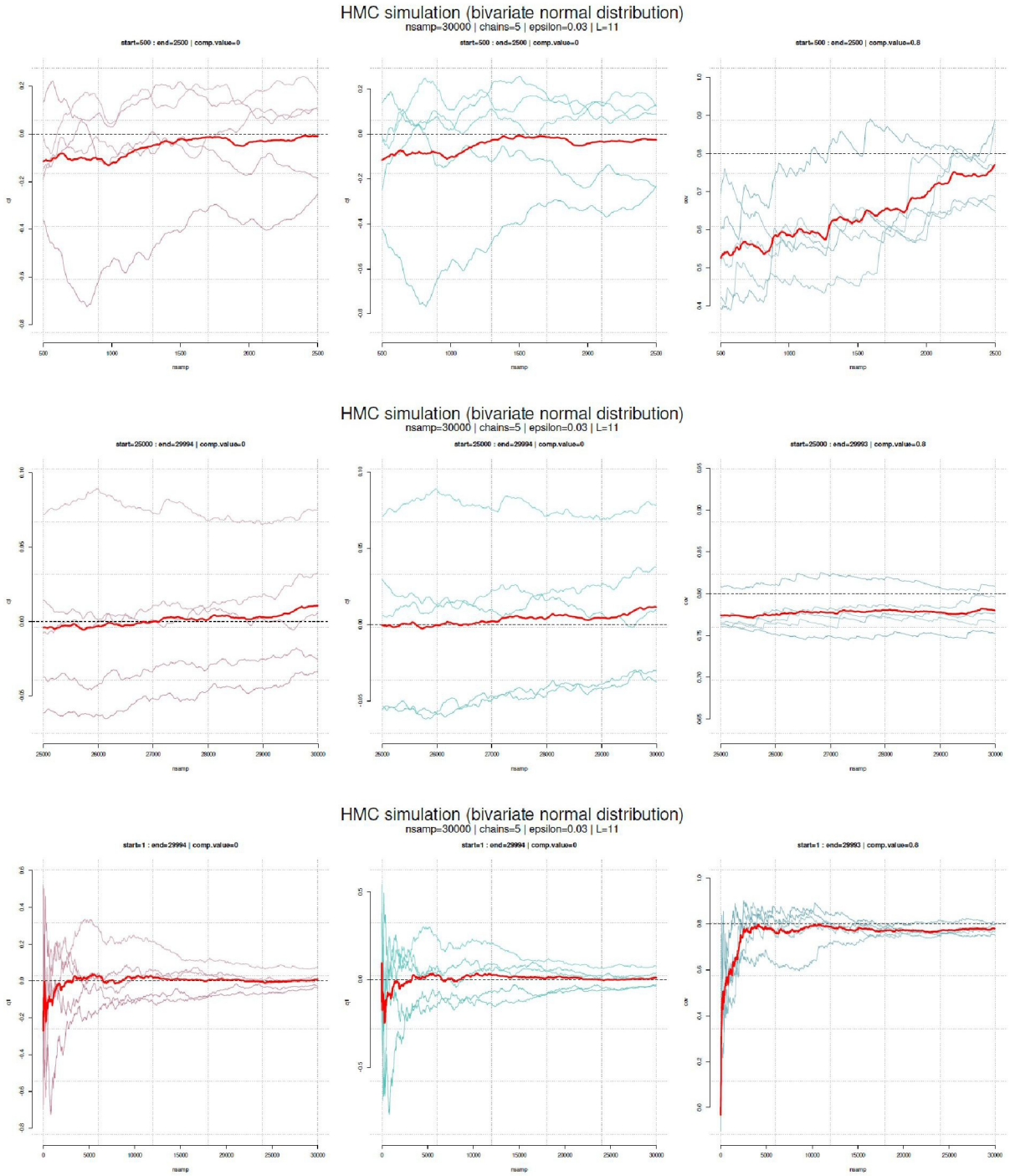


Figura 6.103. *Simulación HMC distribución normal bivariante (evolución de MCMC arriba para [500:2500], centro para [25000:fin] e inferior para [1:fin] para* `q1, q2` *y covarianza,* $\epsilon = 0{,}03$, $L = 11$)

Tarea 6.10: Simulación HMC MCMC y valores de salida

La tarea para los lectores en este punto sería ahora también cambiar el número de leapfrogs L con y sin cambios simultáneos de ϵ. ¿Qué sale? Empezar con pocas repeticiones y muchas cadenas MCMC antes de pasar a cadenas MCMC más largas. Llegar a valores aún mejores es decir, una convergencia aún más rápida de las cadenas MCMC con la misma precisión de los valores medios y de la matriz de covarianza o peor que la que se indica aquí? ¿Se puede compensar una elección errónea del tamaño del paso sólo con el número de leapfrogs?

Escriba un script en R que le permita utilizar menos réplicas (por ejemplo, (p. ej. n = 3000) para recorrer diferentes escenarios de tal forma que obtenga las estadísticas relevantes y obtenga un gráfico final que le ofrezca una visión general resumida.

6.13.2.3.3 HMC comparado con el algoritmo MH – Como contraste final, ahora simulamos la distribución normal bivariante con el algoritmo MH y comparamos el resultado con el de HMC. Tomamos el mismo número de repeticiones, n = 30 000, y observamos la tasa de aceptación y los valores simulados como en el diagnóstico. Primero definimos la función necesaria, ya que en general para el algoritmo HMC la función logarítmica negativa y la función de gradiente no son necesarias, sino sólo la función normal del objetivo de la distribución apuntada.

```
# comparison with Metropolis Hastings R-Code
# no neg log like HMC!
U.MH <- function(q, ...) mvtnorm:::dmvnorm(x=q, mean=c(0,0),
        sigma=sigmamat, log=TRUE)
U.MH(q=c(0.5,0.5))
```

Los valores iniciales del generador aleatorio, el número de repeticiones MCMC y los valores iniciales del algoritmo MH se basan en los del HMC (véase más arriba):

```
set.seed(seeds[1]) R-Code
nsamp <- 3e4
# create vectors for acceptance rate and post values
a.MH <- rep(NA, nsamp)
post.MH <- matrix(NA, nrow=nsamp, ncol=2)
# initial prob
prob.current <- U.MH(q=current.values, sigma=sigmamat)
sd.param <- 1
current.values <- mu
```

y ahora todo el algoritmo MH para una sola ejecución:

```
for (i in 1:nsamp)
{
  #print(current)
  #create proposed values for x and y
  proposed <- c(rnorm(1,current.values[1],sd.param), #x
  rnorm(1,current.values[2],sd.param)) #y
  prop.proposed <- U.MH(q=proposed, sigma=sigmamat)
  H1minusH0 <- prop.proposed-prob.current
  # = min(1,exp(prop.proposed)/exp(prob.current))
  prob.accept <- min(1,exp(H1minusH0))
  testvalue <- runif(1)
```

```
  if(testvalue <= prob.accept)
  {
    current.values <- post.MH[i,] <- proposed
    a.MH[i] <- 1
    prob.current <- prop.proposed
  } else
  {
    #not required
    #post[i,] <- NA
  }
}
```

A diferencia de la HMC , la tasa de no-aceptación es

```
> length(which(is.na(a.MH)))/length(a.MH)
[1] 0.594
```

es decir, 1- 0,594, ~40,6 % de aceptación. No es una cifra especialmente alta. Por otra parte cálculos para este sencillo ejemplo y el procedimiento es mucho más rápido. Esto cambia con modelos complejos, ya que entonces el MH progresa más lentamente que el HMC y puede alcanzar algunas zonas del espacio de parámetros con dificultad – si acaso. Las estadísticas descriptivas para la comparación con la distribución teórica parecen buenas.

```
> # means
> Xbar.MH <- apply(post.MH,2,mean, na.rm=TRUE)
> names(Xbar.MH) <- c("q1","q2")
> Xbar.MH
q1     q2
0.0198 0.0349
> # covariance matrix
> S.MH <- cov(post.MH, use="complete.obs")
> rownames(S.MH) <- colnames(S.MH) <- c("q1","q2")
> S.MH
   q1    q2
q1 1.043 0.813
q2 0.813 1.052
> S.MH/sigmamat
   q1   q2
q1 1.04 1.02
q2 1.02 1.05
> # sd
> apply(post.MH,2,sd, na.rm=TRUE)
q1   q2
1.02 1.03
```

Tanto los valores medios de `q1` y `q2` como la covarianza y las desviaciones estándar se aproximan mucho a los valores teóricos. Por tanto, se ha alcanzado la distribución apuntada. Por interés, comprobemos la distribución generada de este modo con la prueba de Heidelberger-Welch mediante `heidel.diag()` del paquete R `coda`, se repite el problema ya mencionado anteriormente. La primera prueba se ejecuta correctamente, mientras que la segunda genera un error. Esto es de esperar después de lo esperado, ya que aquí también se busca un valor medio de 0.

```
> NA.IDs.MH <- which(is.na(post.MH))
> heidel.diag( post.MH[-NA.IDs.MH] )
Stationarity start p-value
            test iteration
[,1] passed 1     0.774
Halfwidth   Mean  Halfwidth
```

```
     test
[,1] failed 0.0273 0.0359
$eps
[1] 1.36
$itera
[1] 28
$eps.init
[1] 0.01
$steps
[1] 0.05
```

De nuevo, no es sorprendente que la prueba de Heidelberger-Welch rechace la distribución simulada en términos de precisión. Volvemos a añadir 1 a la cadena y la segunda prueba se supera sin problemas.

```
> # repeat wird post.MH + 1
> heidel.diag(post.MH[-NA.IDs.MH] + 1)
Stationarity start p-value
test iteration
[,1] passed 1 0.774
Halfwidth   Mean Halfwidth
test
[,1] passed 1.03 0.0359
```

La figura 6.104 visualiza la distribución resultante con `plotPost()` del paquete R BEST. Contiene el ROPE con 0,5 y 0 como valor de comparación, así como un intervalo de confidencia bayesiana (HDI) del 87 %.

```
# use plotPost from BEST
plotPost(post.MH[-NA.IDs.MH], xlab="quantile",
        ROPE=c(-0.5,0.5), compVal=0, credMass=0.87)
lines(density(post.MH[-NA.IDs.MH]), col="magenta", lwd=3, lty=2)
```

El cálculo de HDI con 87% muestra:

```
> HDInterval::hdi(post.MH[-NA.IDs.MH], credMass=0.87)
lower upper
-1.52 1.57
attr(,"credMass")
[1] 0.87
```

Si se repite todo con n = 5 cadenas MCMC, se obtiene en resumen el siguiente resultado. Para ello, la función de R `bivarsim.MH2()` se diseñó de tal forma que genera cadenas MCMC y las muestrea hasta que se alcance el número de simulaciones apuntado con éxito (es decir, simulaciones aceptadas) y, por tanto, todas las cadenas MCMC tienen la misma longitud. El vector con los porcentajes de aceptación tiende a ser diferente entre las cadenas MCMC, ya que requieren diferentes longitudes de tiempo para alcanzar el objetivo de las aceptaciones establecidas. Se parte del análisis de la tasa de aceptación y el factor necesario para alcanzar el número apuntado de simulaciones con éxito.

```
> # number of NAs
> NAs.anz <- sapply(as, function(x) sum(is.na(x)+0))

> NAs.anz
[1] 44339 44120 44291 44805 44917
> # acceptance rate

> accept.rate.MH <- 1-(NAs.anz/mcmc.MH.l)
> accept.rate.MH
[1] 0.404 0.406 0.407 0.399 0.399
```

```
> summary(accept.rate.MH, sd=sd(accept.rate.MH))
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.399 0.399 0.404 0.403 0.406 0.407

> # factor length extension compared to
> # required number of accepted values
> as.l <- sapply(as, length)
> as.l/nsamp
[1] 2.48 2.47 2.48 2.49 2.50
```

Las cinco cadenas MCMC diferentes tienen tasas de aceptación aproximadamente comparables de 0,4. El factor de extensión es, en consecuencia, de aproximadamente 2,5. Esto significa que para el objetivo de 30 000 repeticiones con éxito, normalmente se necesitan 30 000 * 2,5 = 75 000 repeticiones para alcanzar el objetivo o 25 000 repeticiones para alcanzar el criterio de Kruschke (2016) de ESS = 10 000. Esto debe mejorarse. El análisis recogido mediante el objeto MCMC `coda` utiliza `summary()` para mostrar el resumen estadístico a través de las cadenas MCMC. No hay nada de lo que quejarse aquí. Tanto la media de 0 y la desviación estándar de 1 están suficientemente bien aproximadas. Este muestra el gráfico bivariante con elipses (véase la Fig. 6.105).

```
> summary(OUTmcmc.list)
Iterations = 1:30000
Thinning interval = 1
Number of chains = 5
Sample size per chain = 30000

1. Empirical mean and standard deviation for each variable,
        plus standard error of the mean:
   Mean    SD   Naive SE Time-series SE
q1 0.00549 1.01 0.00261  0.00726
q2 0.00581 1.01 0.00260  0.00718

2. Quantiles for each variable:
    2.5%  25%   50%      75%   97.5%
q1 -1.97 -0.683 0.000672 0.689 1.98
q2 -1.98 -0.680 0.000859 0.690 1.97
```

La figura 6.106 muestra los trazados asociados, las estimaciones de densidad y los trazados de violín de las distintas cadenas MCMC. La figura 6.107 muestra a su vez la evolución de las cadenas MCMC. Obsérvese que dada una tasa de aceptación media de ~0,4, hay significativamente menos repeticiones en el gráfico que repeticiones *con éxito y sin éxito* aparecían por cadena MCMC. Esto demuestra la superioridad del HMC sobre el algoritmo MH.

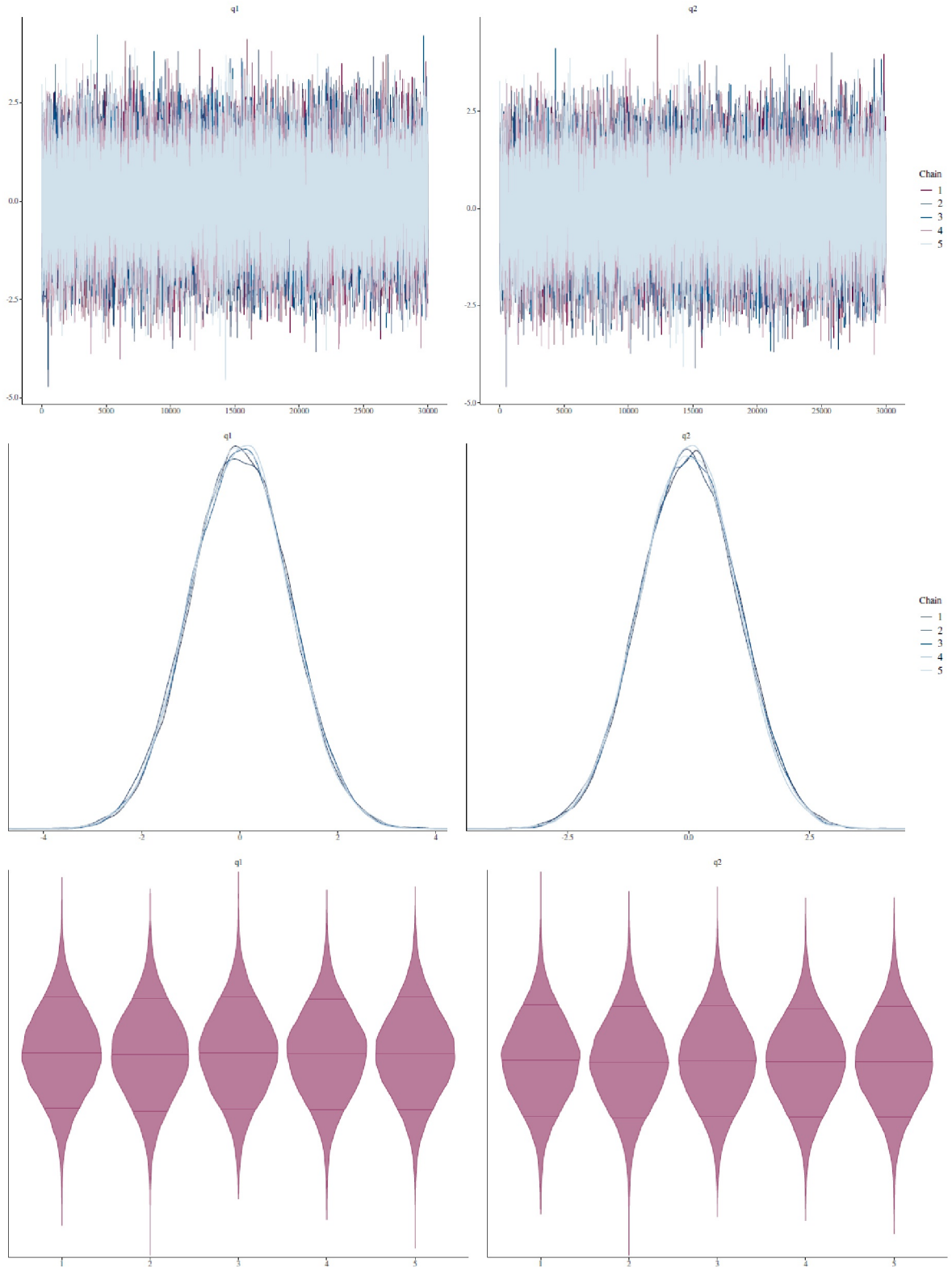


Figura 6.106. *Distribución normal bivariante de la simulación MH*
(trazado, trazado de densidad, trazado de violín para q1 y q2, 30 000 repeticiones, 5 cadenas MCMC)

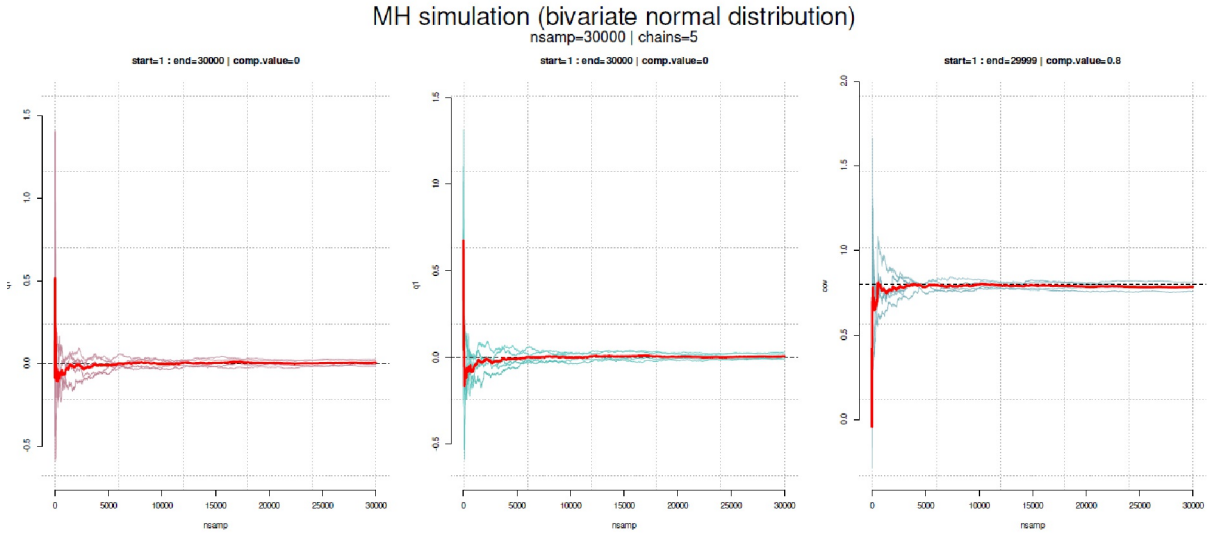


Figura 6.107. *Distribución normal bivariante de la simulación MH*
(evolución de las cadenas MCMC para q1 y q2, 30 000 repeticiones, 5 cadenas MCMC).

6.13.3 Diagnóstico de cadenas MCMC

Los diagnósticos MCMC suelen centrarse en la convergencia de la cadena MCMC y en si existe un proceso estocástico estacionario. Existen muchas analogías con los procesos físicos de la teoría de detección de señales y, en general, con el manejo numérico de señales (frecuencias y amplitudes) y sus distribuciones. Por lo tanto, en primer lugar nos referiremos a algunos de estos términos y conceptos, ya que la bibliografía pertinente hace referencia a ellos.

Se habla de proceso estacionario cuando la distribución de un proceso estocástico no depende de su desplazamiento en el tiempo. Parámetros como la media y la varianza no cambian con el tiempo. Físicamente, la frecuencia y la amplitud serían constantes. Si ahora actúa una fuerza externa sobre este proceso estacionario, pierde su estacionariedad, ya que entonces cambiarían la frecuencia y/o la amplitud y, en consecuencia, los estadísticos que dependen de ellas. Para los procesos o series temporales con tendencia o ciclo, la estacionariedad no se da – para los procesos sin tales características, sin embargo, sí. Por lo tanto, una de las principales preocupaciones de los análisis de series temporales es transformar los procesos no estacionarios en estacionarios. Esto puede hacerse, por ejemplo, eliminando tendencias, periodicidades, etc., que en su mayoría sólo dependen del tiempo, es decir, de patrones en la secuencia de los datos. Lo que queda después de estos análisis es un proceso estacionario, pero esto debe comprobarse por separado en el caso concreto. Para seguir tratando procesos estacionarios, naturalmente se intenta sacar conclusiones sobre la evolución futura a partir de las características del pasado, ya que no se puede predecir el futuro. Las características estadísticas para la predicción, como el valor medio, la varianza, la autocorrelación, etc., sirven a este objetivo, ya que se supone que permanecen constantes a lo largo del tiempo en los procesos estacionarios. La aplicación de modelos de series temporales y otros métodos de análisis a datos secuenciales persigue el único objetivo de transformarlos en un proceso estacionario y describir o analizar la serie en su conjunto. En relación con las cadenas MCMC, la convergencia o estacionariedad es el requisito previo para todos los análisis posteriores, es decir, un componente esencial.

Un análisis espectral proporciona información sobre el contenido de las frecuencias y las fuentes de variación en series temporales y en cadenas MCMC. La representación espectral de un proceso estocástico estacionario es un cuasi-equivalente a la expansión de la serie de Fourier y tiene como objetivo averiguar los procesos y estructuras subyacentes de la serie. En física, la densidad espectral describe a su vez el contenido de frecuencias de una señal, es decir, la fuente de una serie de datos, que es la distribución apuntada en las cadenas MCMC. La densidad espectral es una función de frecuencia y no de tiempo. La función de densidad espectral indica la intensidad con la que se produce una frecuencia en un espectro de interés. Puede entenderse como la distribución de energía donde y con qué intensidad la forman los componentes respectivos de una señal.

Las cadenas MCMC, como expresión de las probabilidades posteriores en la estadística de Bayes, también tienen diferentes intensidades para diferentes rangos de valores, es decir, probabilidad para un determinado valor de la distribución de parámetros. Como puede verse, la analogía con la física resulta ser pertinente, ya que tanto la terminología como la tecnología utilizadas se adaptan muy bien al análisis de las cadenas MCMC. El objetivo de la estimación de la densidad espectral es, pues, estimar la densidad espectral de una señal aleatoria a partir de una serie de muestras a lo largo del tiempo.

Un caso especial (extremo) es el ruido blanco, análogo a la luz blanca. En este caso, todas las frecuencias están representadas por igual y la densidad espectral es constante. Por lo tanto, no hay más estructura en los datos. Este sería el caso ideal de un proceso estacionario. La densidad espectral sólo depende de la varianza σ^2. Por tanto, si la frecuencia es λ, la densidad espectral *f(λ)* del ruido blanco es

$$\mathrm{f}(\lambda) = \frac{1}{2 \cdot \pi} \cdot \sigma^2 \tag{6.146}$$

y con la función de covarianza γ*(h)* asociada

$$f(\lambda) = \begin{cases} \sigma^2 & \text{wenn} h = 0 \\ 0 & \text{sonst} \end{cases} \tag{6.147}$$

Una fórmula general de la densidad espectral *f(λ)* para el proceso estocástico estacionario X_t es la siguiente

$$f(\lambda) = \frac{1}{2 \cdot \pi} \sum_{h=-\infty}^{\infty} e^{-i \cdot h \cdot \lambda} \cdot \gamma(h) \tag{6.148}$$

con la representación espectral γ*(h)* y *f* como la transformada de Fourier de la función de covarianza γ, que, debido a la estacionariedad, sólo depende de la diferencia temporal *h*. Para γ*(h)* como función de covarianza en la diferencia temporal *h* se aplica lo siguiente

$$\mathrm{Cov}(X_t, X_{t+h}) = \Gamma(t, t+h) =: \gamma(h) \tag{6.149}$$

Existen varias posibilidades técnicas para estimar la densidad espectral, tanto no paramétricas como paramétricas. Entre los métodos paramétricos que interesan aquí figuran los modelos autorregresivos (= AR), que suelen ser adecuados para el análisis de series temporales. Un análisis de series temporales es, por tanto, un análisis estadístico de puntos de datos sucesivos, generalmente medidos a lo largo del tiempo, con el fin de averiguar las características específicas de una serie temporal. Los modelos AR, por su parte, examinan las autocorrelaciones dentro de una serie (temporal). Los modelos AR ofrecen una alternativa a la transformación discreta de Fourier (= análisis espectral).

En el modelo AR, cada valor de la serie temporal se estima por regresión sobre su valor anterior. El número de valores anteriores (pasados) se denomina *orden del modelo*. Además de los valores medios y los picos de una distribución, la descomposición de una señal en sus componentes, especialmente en series temporales, va seguida del análisis de la varianza – en el caso de MCMC, la de la distribución apuntada –

a lo largo del tiempo. La varianza como potencia media de un proceso puede obtenerse integrando el espectro de potencia en todas las frecuencias.

Para el análisis de las cadenas MCMC, se examina si las muestras de la cadena MCMC son independientes entre sí o no, lo que responde a la cuestión de la autocorrelación y si (aún) siguen una tendencia o una periodicidad, lo que hablaría de un no-éxito de la simulación MCMC. Entonces, los datos sucesivos dependerían en gran medida de los valores de los anteriores y no del proceso estocástico subyacente. En los diagnósticos numéricos y gráficos de las cadenas MCMC, especialmente el valor medio, las varianzas mencionadas (dentro de las cadenas MCMC y entre ellas), así como las autocorrelaciones dentro de las cadenas MCMC, desempeñan un papel importante, además de la densidad espectral mencionada. La densidad espectral a frecuencia 0 desempeña un papel aparte. En física, la frecuencia 0 significa un término constante en el que no se produce ningún movimiento (onda, pico, etc.).

En este punto, por ejemplo, una onda tendría un periodo y una longitud de onda infinitos y el tiempo entre picos también sería infinito. La estimación de la densidad espectral a frecuencia 0 en los modelos AR tiene por objeto determinar la varianza del valor medio de una serie temporal para obtener el error típico de la media. A continuación, se incluyen la varianza y el error típico en los cálculos para comprobar partes de cadenas MCMC entre sí, comparar cadenas MCMC entre sí, etc. La función `spectrum0.ar()` del paquete `coda` de R calcula la densidad espectral en frecuencia cero utilizando un modelo AR de la siguiente manera:

```
v0[i] <- ar.out$var.pred/(1 - sum(ar.out$ar))^2
```

En el extracto anterior del código R de `spectrum0.ar()`, `ar.out` es el R- objeto con el modelo AR estimado, `var.pred` es la varianza de predicción, es decir, una estimación de la porción de varianza de la serie temporal que *no puede ser explicada* por el modelo AR, es decir, una varianza residual. Además, `ar` son los coeficientes autorregresivos estimados del modelo ajustado. La fórmula establece que la varianza *no explicada* por el modelo AR se divide por el cuadrado de la diferencia de los coeficientes AR sumados a 1. En resumen, los diagnósticos MCMC comunes examinan así las varianzas y los valores medios de una forma muy clásica, pero relacionada con las series temporales. Se intenta identificar la fase de burn-in y, posteriormente, dividirla, y se comprueba la convergencia de todo el proceso. También se plantea la cuestión de la precisión, es decir, la exactitud de la estimación – todo ello en el contexto de las autocorrelaciones potencialmente existentes en las series temporales, que suelen reducir la precisión.

Los métodos de diagnóstico MCMC más conocidos son

- Diagrama de Gelman-Rubin-Brooks y diagnósticos de Gelman-Rubin (véase el capítulo 6.13.3.1).
- Diagnóstico de Geweke (véase el capítulo 6.13.3.2)
- Diagnóstico de Heidelberger-Welch (véase el capítulo 6.13.3.3)
- Diagnóstico de Raftery-Lewis (véase el capítulo 6.13.3.4)
- Análisis de autocorrelaciones mediante gráficos (véase el capítulo 6.13.3.5)
- Tamaño efectivo de la muestra (véase el capítulo 6.13.3.6)

El gráfico Gelman-Rubin-Brooks y los estadísticos comparables (véase el capítulo 6.6.1) son adecuados para comparar cadenas MCMC entre sí. Utilizaremos el caso de ejemplo anterior con $\epsilon = 0{,}1$, cinco cadenas MCMC y 30.000 réplicas y comenzaremos con el gráfico Gelman-Rubin-Brooks (Brooks & Gelman, 1998) y el diagnóstico Gelman-Rubin (Gelman & Rubin, 1992).

6.13.3.1 Gráfico de Gelman-Rubin-Brooks y diagnóstico de Gelman-Rubin

El diagnóstico de Gelman-Rubin se basa en la idea de comparación de varianzas y "olvido". Las cadenas MCMC se comparan en paralelo y, por tanto, el diagnóstico no puede calcularse para cadenas MCMC individuales. La estacionariedad se alcanza cuando las cadenas MCMC han olvidado cuáles eran sus valores iniciales y se comportan de forma independiente. Si esto se considera para todas las cadenas MCMC, en este punto se habrán mezclado indistintamente. Se considera que los valores iniciales están sobre-dispersados en comparación con la distribución posterior. La sobre-dispersión se da cuando la varianza empírica s^2 en los datos es significativamente mayor que la varianza teórica σ^2 esperada por el modelo, es decir, $s^2 \gg \sigma^2$.

Gráficamente, se puede investigar esto superponiendo trazados, lo que se hace mediante `mcmc_trace()` del paquete `bayesplot` de R. Ninguna de las cadenas MCMC debería destacarse o desviarse de las demás cuando se superponen. Idealmente, parece un batiburrillo sin tendencias, periodos, etc. alrededor del valor medio.

Numéricamente, la idea de olvido puede implementarse en los diagnósticos de Gelman-Rubin de esta forma: la varianza *dentro* de las cadenas MCMC se compara con la varianza *entre* las cadenas MCMC. cadenas. Esto corresponde a la idea del análisis clásico de la varianza. La varianza se puede calcular de la distribución estacionaria como el valor medio de las varianzas empíricas *W* dentro de cada cadena MCMC o como la de las cadenas MCMC combinadas entre sí $\hat{\sigma}^2$. Partiendo de $m > 1$ cadenas MCMC, n como el número de réplicas por cadena MCMC y la entre-varianza empírica B/n (= entre las cadenas MCMC), $\hat{\sigma}^2$ viene dada por:

$$\hat{\sigma}^2 = \frac{(n-1)\cdot W}{n} + \frac{B}{n} \tag{6.150}$$

En el caso de convergencia ambas estimaciones son insesgadas. De lo contrario, el primer método de varianza interna subestimará la varianza porque las cadenas MCMC individuales aún no han pasado por todo el espacio de parámetros y, por tanto, aún no se ha capturado toda la varianza. En el segundo caso de entre-varianza, se produce una sobre-estimación de la varianza, ya que los valores iniciales están sobredispersos. El diagnóstico de convergencia *R* se basa en *W* (= varianza interna), la varianza combinada $\hat{V}$ a través de las cadenas MCMC (= entre-varianza) y los grados de libertad $\hat{d}$, estimados mediante el método de los momentos y basados en la distribución *t*, ya que se trata de valores estimados. El diagnóstico *R* también se denomina PSRF (= potential scale reduction factor). Además, $\hat{\mu}$ es la media muestral estimada de todas las cadenas MCMC, $\hat{V}$ es la varianza combinada estimada asociada y *d* los grados de libertad distribuidos según *t*. A continuación:

$$\hat{V} = \hat{\sigma}^2 + \frac{B}{m\cdot n} \tag{6.151}$$

$$d = \frac{2\cdot \hat{V}^2}{\text{var}(\hat{V})} \tag{6.152}$$

$$R = \sqrt{\frac{(d+3)\cdot \hat{V}}{(d+1)\cdot W}} \tag{6.153}$$

Una diferencia del procedimiento para el caso multivariante en comparación con el caso univariante es que la versión multivariante no incluye un ajuste para los grados de libertad estimados.

Una regla empírica dice que el diagnóstico Gelman-Rubin (= factor de reducción de escala para cada parámetro) está bien siempre que sea inferior a 1,1. Un factor de 1 significa que la entre-varianza y la dentro-varianza de las cadenas MCMC son iguales y no hay diferencias significativas. Los valores superiores a 1 representan diferencias y ausencia de convergencia. Dado que se trata de un intervalo creíble bayesiano (= IC) creado con la distribución *t*, éstos son demasiado amplios en caso de que falle la convergencia. Sin

embargo, los IC tienen el potencial de encogerse por el llamado Shrink-factor (Factor de encogimiento) si la serie MCMC continúa y no hay problemas significativos en el propio modelo.

El gráfico Gelman-Rubin-Brooks muestra el desarrollo de la reducción de escala a lo largo del tiempo. Los burn-ins deben ignorarse, ya que aquí se pueden producir naturalmente diferencias en función de los valores iniciales. Es importante examinar si más adelante en el transcurso el factor vuelve a ser mayor que 1 o si permanece relativamente estable (`ptII_quan_Bayes_HMC.r`).

```
> gelman.plot(OUTmcmc.list)
> gelman.diag(OUTmcmc.list)
Potential scale reduction factors:
    Point est. Upper C.I.
q1  1          1
q2  1          1
Multivariate psrf
1
```

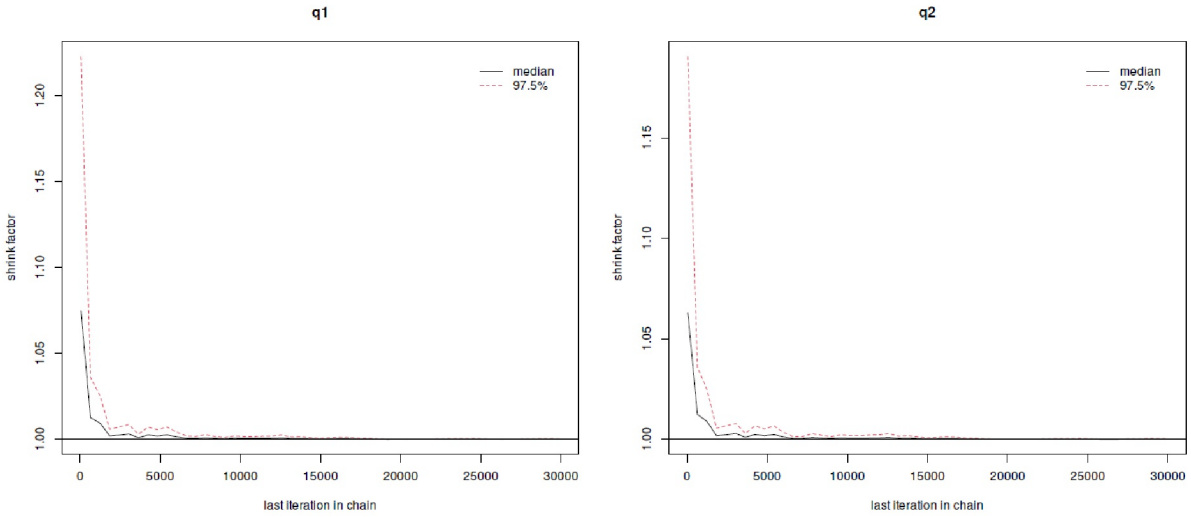


Figura 6.108. *Simulación HMC distribución normal bivariante*
(MCMC de diagnóstico, diagrama de Gelman, = 0,1, L = 11)

Ni en la salida numérica ni en el gráfico de Gelman-Rubin-Brooks (véase la Fig. 6.108) hay irregularidades – la estacionariedad y la convergencia están dadas desde este punto de vista.

6.13.3.2 El diagnóstico de Geweke

El diagnóstico de Geweke (Geweke, 1992) puede utilizarse para determinar con más detalle el periodo de burn-in, es decir, la parte inicial más pequeña de la cadena MCMC que supera el diagnóstico. Se basa en series temporales y equivale aproximadamente a comparar las medias de dos muestras comparando el primer 10% con el último 50%. Siempre que las muestras se hayan extraído de la distribución estacionaria de la cadena MCMC, las dos medias son iguales y el estadístico de Geweke tiene una distribución normal estándar asintótica. Así pues, la H_0 es que la cadena MCMC se encuentra en la distribución estacionaria y crea estadísticos *z* para cada parámetro estimado. La prueba se realiza de forma clásica-estadística y se basa en el no rechazo de la H_0. En el caso sencillo, el diagnóstico de Geweke se crea a partir de las dos partes A y B de la cadena MCMC θ y longitud *n*:

$$Z_n = \frac{\bar{\theta}_A - \bar{\theta}_B}{\sqrt{\frac{1}{n_A}\hat{S}_\theta^A(0) + \frac{1}{n_B}\hat{S}_\theta^B(0)}} \qquad (6.154)$$

La estadística de la prueba es, por tanto, un valor *z* estándar, creado a partir de la diferencia entre las dos medias dividida por el error estándar estimado. Este último se mide mediante la densidad espectral S^(0) a frecuencia 0 y tiene en cuenta las autocorrelaciones.

Dado que el diagnóstico de Geweke sólo está pensado para el caso univariante, se necesita una ampliación para el caso bivariante a fin de tener en cuenta las covarianzas: en el caso multivariante, se necesitan diagnósticos adicionales para el tratamiento adecuado de las múltiples covarianzas entre las variables. En el caso bivariante de *X* e *Y*, $y_i = (xi - \bar{X})*(y_i - \bar{Y})$ se calcula para cada *X* e *Y* y, a continuación, se aplican los diagnósticos de Geweke a la muestra de *y* (`ptII_quan_Bayes_HMC.r`).

```
> geweke.OUTmcmc.list.z <- do.call("rbind",
        + lapply(geweke.diag(OUTmcmc.list), function(x) x$z))
> geweke.OUTmcmc.list.frac <- do.call("rbind",
        + lapply(geweke.diag(OUTmcmc.list), function(x) x$frac))
> colnames(geweke.OUTmcmc.list.frac) <- c("frac 1st", "frac 2nd")
> cbind(geweke.OUTmcmc.list.z, geweke.OUTmcmc.list.frac)
      q1     q2    frac 1st frac 2nd
[1,]  1.163  1.000    0.1      0.5
[2,]  0.829  0.853    0.1      0.5
[3,]  0.670  0.724    0.1      0.5
[4,] -1.563 -1.699    0.1      0.5
[5,] -1.253 -1.132    0.1      0.5
```

Para `q1` y `q2`, se obtienen valores *z*. Una interpretación estadística clásica es que si los valores *p* de las pruebas *z* son superiores a un umbral crítico, como el incalificable 5 %, es decir, "no significativo" en el lenguaje de la estadística clásica, entonces la cadena MCMC ha convergido. Entonces, el primer 10 % de la cadena MCMC procede de la misma distribución que el último 50 % y no hay mayores problemas. La prueba es contra H_0, una circunstancia a la que la estadística clásica sólo puede hacer frente – si es que puede – con tirones como en las pruebas de equivalencia (véase el capítulo 4.4.9). El no-rechazo de H_0 no conduce a ninguna ganancia real de conocimiento, ya que no se trata de una prueba directa de la validez de H_0. Así pues, cabe preguntarse por qué utilizar una prueba estadística clásica en un contexto bayesiano, pero no pasa nada por intentarlo:

```
> apply(geweke.OUTmcmc.list[,c(1:2)],2,
        + function(x) pnorm(abs(x), lower.tail=FALSE))*2
     q1    q2
[1,] 0.245 0.3171
[2,] 0.407 0.3937
[3,] 0.503 0.4691
[4,] 0.118 0.0894
[5,] 0.210 0.2575
```

Como se puede ver, la interpretación no contradice el supuesto de igualdad de medias. Se puede representarlo gráficamente con `geweke.plot()`, pero no se imprime aquí, sino sólo los valores *z*.

6.13.3.3 Diagnóstico de Heidelberg-Welch

El diagnóstico de Heidelberger-Welch (Heidelberger & Welch, 1981, 1983) usa la estadística Cramervon-Mises para probar la hipótesis nula de que los valores de simulación generados se originan de una distribución estacionaria. En primer lugar, la prueba se aplica a toda la cadena MCMC y, a continuación,

tras eliminar el primer 10%, 20%, ... de los datos – hasta que o bien la hipótesis nula no se pueda rechazar o se haya eliminado el 50% de la cadena MCMC. Esto último produce un fallo con respecto a la prueba de estacionariedad e indica que se necesita una cadena MCMC más larga o que el modelo no es correcto. Si la prueba de estacionariedad tiene éxito `heidel.diag()` informa del número de iteraciones pertenecientes a la fase de burn-in y del valor *p* clásico de éxito en la *prueba S-* (prueba de convergencia) o *prueba H* (prueba de precisión de la media) y los valores asociados (media, valores del intervalo medio probado).

Tras la prueba global de estacionariedad, se examina la precisión de la medición en una segunda parte de la prueba. Se examina la parte de los datos que se considera no quemada y, por tanto, estacionaria. Estos datos se tratan como una serie temporal a partir de la cual se estima la densidad espectral *S(0)* a frecuencia se estima 0. El error típico asintótico de la media se obtiene como en el diagnóstico de Geweke (Geweke, 1992) a partir de

$$SE_{\bar{X}} = \frac{S(0)}{N_p} \tag{6.155}$$

N_p corresponde a la longitud de la cadena MCMC estacionaria restante. A continuación se calcula un intervalo de confianza del 95 % en torno al valor medio $\bar{x}$ utilizando el error típico asintótico $SE_{\bar{x}}$ y la mitad de este intervalo se compara con el valor medio. La cuestión es si el cociente de la mitad de la anchura del intervalo dividido por la media es menor que ϵ, es decir, si

$$\frac{0.5 \cdot CI_{\bar{X}(95\%)}}{\bar{X}} < \epsilon \tag{6.156}$$

Si es así, se considera que se ha superado la prueba. ϵ se selecciona como fracción pequeña, ya que es una cuestión de precisión. En el paquete `coda` de R, la función `heidel.diag()` tiene un valor por defecto de $\epsilon = 0.1$. Si el cociente es mayor que ϵ, la prueba se considera fallida y se requiere una cadena MCMC más larga para superarla. La cadena MCMC es necesaria para obtener la precisión requerida del parámetro de interés, suponiendo que la estacionariedad ya se ha confirmado con éxito de antemano (`ptII_quan_Bayes_HMC.r`).

```
> heidelbergwelch.OUTmcmc.list <- cbind(
        + chain=rep(1:nchains,each=2),
        + do.call("rbind",
        + lapply(heidel.diag(OUTmcmc.list), function(x) x[1:2,1:6])
        + ))
> heidelbergwelch.OUTmcmc.list
   chain stest start pvalue htest mean    halfwidth
q1 1     1     1     0.0546 0     0.02258 0.0247
q2 1     1     1     0.0793 0     0.02299 0.0244

q1 2     1     1     0.5984 0     0.00951 0.0252
q2 2     1     1     0.6730 0     0.01048 0.0250

q1 3     1     1     0.4252 0     0.00172 0.0243
q2 3     1     1     0.2580 0     0.00239 0.0247

q1 4     1     1     0.1721 0    -0.00794 0.0248
q2 4     1     1     0.1564 0    -0.00969 0.0248

q1 5     1     1     0.2650 0    -0.00922 0.0248
q2 5     1     1     0.3057 0    -0.00884 0.0245
```

En la salida R se observa en todo momento que la primera prueba de estacionariedad es positiva (= 1) y la segunda prueba de precisión es negativa (= 0). Esto significa que, desde el punto de vista de la prueba de Heidelberg-Welch, las cadenas MCMC son todas estacionarias, pero el valor medio no se determina con

suficiente precisión. Merece la pena un examen más detallado y empezamos por ϵ: Para entender cómo tendría que cambiar ϵ para que la prueba tuviera éxito, la función R `heidel.eps.det()` comienza con ϵ = 0.1 y aumenta en una cierta cantidad por pasos de iteración hasta que la prueba es positiva. Elegimos la mitad de ϵ, es decir, *0.5*ϵ, como el tamaño del paso y establecemos `steps=0.05` en consecuencia. `heidel.eps.det()` da como resultado el ϵ-valor mínimo para que la prueba de la mitad pase con éxito e `itera` indica el número de iteraciones hasta que el aumento por pasos de ϵ tuvo éxito. Esto da una impresión aproximada en qué órdenes de magnitud termina el problema y se puede entonces comparar la norma de ϵ = *0.1* con el valor determinado empíricamente. Después, vale la pena considerar cómo puede producirse esta situación en primer lugar, es decir, que se haya perdido la precisión.

```
eps.det.OUTmcmc.list.q1 <- data.frame(t(sapply(1:nchains,
function(x) heidel.eps.det(OUTmcmc.list[[x]][,"q1"]))))
eps.det.OUTmcmc.list.q2 <- data.frame(t(sapply(1:nchains,
function(x) heidel.eps.det(OUTmcmc.list[[x]][,"q1"]))))
```

Los resultados de `q1` y `q2` son los siguientes: dado que se estima una distribución normal bivariante con las mismas medias y desviaciones estándar, no es sorprendente que los valores de `q1` y `q2` sean idénticos. Sin embargo, *no lo son* para las cadenas MCMC individuales:

```
> eps.det.OUTmcmc.list.q1
    eps   itera  eps.init steps
1    1.11  23    0.01     0.05
2    2.66  54    0.01     0.05
3   14.1  283    0.01     0.05
4    3.16  64    0.01     0.05
5    2.71  55    0.01     0.05
> eps.det.OUTmcmc.list.q2

    eps   itera  eps.init steps
1    1.11  23    0.01     0.05
2    2.66  54    0.01     0.05
3   14.1  283    0.01     0.05
4    3.16  64    0.01     0.05
5    2.71  55    0.01     0.05
```

Obviamente, las cadenas MCMC difieren a pesar de su longitud de 30 000 repeticiones para el caso simple de la distribución normal bivariante. Ahora bien, se podría pensar que el ejemplo de la distribución normal bivariante con 30 000 repeticiones por cadena MCMC y una tasa de no-aceptación muy modesta debería conducir a una precisión suficiente. Si se cuestiona la precisión, merece la pena dar un paso atrás en el funcionamiento de la prueba de Heidelberg-Welch y comprobar qué es lo que se cuestiona aquí. En el presente ejemplo se pretende una media de 0 y una desviación estándar de 1 más la covarianza ϱ. Los valores simulados ya se aproximan a 0, como deja claro la salida resumida. En consecuencia, la comparación necesaria de la mitad de la anchura del intervalo de confianza del 95 % dividida por la media cercana a 0 y, a continuación, la comparación a ϵ = *0.1* es tal que, independientemente de lo estrecho que sea el intervalo de confianza, el resultado final se divide por un valor próximo a 0 y el cociente se hace globalmente muy grande y ciertamente no es menor que ϵ. Si se observa un ejemplo de cálculo independiente del caso

```
> # non-empirical example
> CI.width <- 0.1
> mean.emp <- 0.003
> epsilon <- 0.1
> ratio.CI.mean <- 0.5*CI.width/mean.emp
> ratio.CI.mean
[1] 16.7
> ratio.CI.mean < epsilon
[1] FALSE
```

se nota inmediatamente lo grande que es el cociente de la mitad del intervalo dividido por la media empírica en comparación con ϵ. Obviamente, es prácticamente imposible pasar la prueba de esta manera. Por lo tanto, no tiene mucho sentido realizarla, porque es difícil calcular con 0 e infinito. En consecuencia, la prueba de Heidelberger-Welch sólo es adecuada para medir la precisión sólo si los valores se encuentran fuera de 0. Esta es la solución para el presente caso y `heidel.diag()` debería dar un aviso para valores cercanos a 0.

Ilustraremos este proceso de pensamiento como una breve excursión con la simulación de una simple distribución normal estándar. Fijamos generosamente el número de extracciones en 1 millón.

```
> # simulate simple normal distribution
> set.seed(seeds[1])
> nd.data <- rnorm(n=1e6)
> summary(nd.data)
Min.  1st Qu. Median Mean 3rd Qu. Max.
-4.37 -0.67   0.00   0.00 0.67    4.96
> sd(nd.data)
[1] 1
> heidel.diag(nd.data)
Stationarity start  p-value
     test iteration
[,1] passed 1       0.967
Halfwidth Mean Halfwidth
     test
[,1] failed -0.000557 0.00196
> eps.det.nd.data <- heidel.eps.det(nd.data)
> eps.det.nd.data
Stationarity start    p-value
     test   iteration
[,1] passed 1         0.967
Halfwidth Mean Halfwidth
     test
[,1] passed -0.000557 0.00196
> eps.det.nd.data
$eps
[1] 3.56
$itera
[1] 72
$eps.init
[1] 0.01
$steps
[1] 0.05
```

A partir de las expectativas anteriores, la cadena de simulación falla la segunda prueba con $\epsilon = 0.1$. No hay duda de que si se extraen 1 000 000 de números aleatorios directamente de la distribución apuntado, se obtienen resultados correctos. Si, después de decenas de extracciones, todavía no se alcanza la precisión desde el punto de vista de la prueba, pero es indiscutible que se ha alcanzado la distribución apuntado, debe haber otra razón. En la Fig. 6.109, las distribuciones teórica y simulada son prácticamente indistinguibles. La precisión de la medición es sin duda exacta y dada:

```
# plot
hist(nd.data,prob=T, pre.plot=grid(), ylim=c(0,0.4),
main="Simulation normal distribution")
lines(density(nd.data), col="darkred")
sek <- seq(-3,3,0.01)
lines(sek,dnorm(sek),col="blue")
```

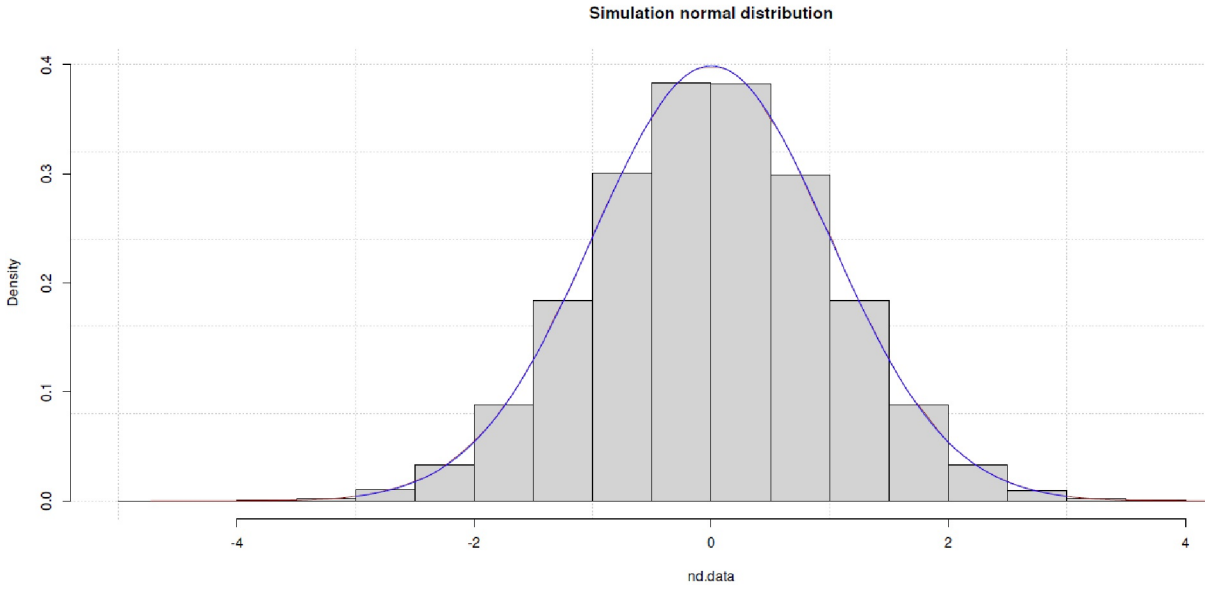


Figura 6.109. *Prueba de Heidelberg-Welch (simulación de distribución normal, comparación con valores aleatorios NV)*

Apliquemos la idea de valores alejados de 0 simplemente añadiendo 1 al conjunto de datos `nd.data`. Una repetición de la prueba de Heidelberger-Welch muestra ahora un resultado diferente:

```
> # repeat wird ndata+1
> heidel.diag(nd.data+1)
Stationarity start    p-value
     test   iteration
[,1] passed 1         0.967
            Halfwidth Mean Halfwidth
     test
[,1] passed 0.999          0.00196
```

Ahora se supera inmediatamente la segunda prueba y se considera que la precisión está asegurada. Esto debería hacer reflexionar sobre si una prueba estadística clásica es adecuada en este caso. Se podría argumentar que una media de casi 0 no se da en la realidad. Sin embargo, las pruebas de equivalencia (véase el capítulo 4.4.9) cuentan exactamente lo contrario. La detección de diferencias inexistentes es tan relevante como la constatación de diferencias. Aplicada a la distribución normal bivariante simulada, se añade 1 a todos los valores

```
# repeat and apply on original data with +1
OUTmcmc.nonas.onlyqs.plus1 <- lapply(OUTmcmc.nonas.onlyqs,
        function(x) x+1)
OUTmcmc.list.plus1 <- as.mcmc.list(lapply(OUTmcmc.nonas.onlyqs.plus1,
        mcmc))
heidelbergwelch.OUTmcmc.list.plus1 <- cbind(chain=rep(1:nchains,
        each=2), do.call("rbind",
        lapply(heidel.diag(OUTmcmc.list.plus1),
        function(x) x[1:2,1:6])))
```

y el resultado de la prueba Heidelberger-Welch cambia de esta manera:

```
> heidelbergwelch.OUTmcmc.list.plus1
   chain stest start pvalue htest mean  halfwidth
q1 1     1     1     0.0546 1     1.023 0.0247
q2 1     1     1     0.0793 1     1.023 0.0244

q1 2     1     1     0.5984 1     1.010 0.0252
q2 2     1     1     0.6730 1     1.010 0.0250

q1 3     1     1     0.4252 1     1.002 0.0243
q2 3     1     1     0.2580 1     1.002 0.0247

q1 4     1     1     0.1721 1     0.992 0.0248
q2 4     1     1     0.1564 1     0.990 0.0248

q1 5     1     1     0.2650 1     0.991 0.0248
q2 5     1     1     0.3057 1     0.991 0.0245
```

Ahora muestra el htest pasado para la precisión para todas las cadenas MCMC sin ningún problema – lo que debía mostrarse.

6.13.3.4 Diagnóstico de Raftery-Lewis

El diagnóstico de Raftery-Lewis (Raftery y Lewis, 1992, 1995) se centra en la precisión de la medición de los cuantiles. En realidad, está diseñado para aplicarse a una sección piloto corta de la cadena MCMC. Examina cuántas iteraciones son necesarias para estimar el cuantil q dentro de una precisión de $\pm r$ con probabilidad s. Se realizan cálculos separados por variable y por cadena MCMC. Si el número de iteraciones es demasiado pequeño, aparece un mensaje de error. La longitud mínima de la muestra MCMC necesaria de la cadena MCMC se da si no existen correlaciones entre las muestras subsiguientes. El índice I (= factor de dependencia) indica hasta qué punto las autocorrelaciones distorsionan el tamaño de muestra requerido de forma racional. El factor de dependencia I representa el aumento proporcional de iteraciones debido a las dependencias secuenciales dentro de la cadena MCMC. La fórmula para I es ahora

$$I = \frac{N}{N_{\min}} \tag{6.157}$$

donde N es el número total de iteraciones que deben ejecutarse para cada variable y N_{min} el número mínimo de iteraciones necesarias para estimar el cuantil específico con la precisión deseada, suponiendo que todas las muestras son independientes entre sí. Esto hace que N sea una cantidad teórica basada en la varianza binomial y, por tanto, proporciona un límite inferior de la cadena MCMC mínima. Al aumentar la probabilidad s y mayor precisión r, N_{min} también crece. El cálculo de N_{min} es

$$N_{\min} = \left[\frac{1}{\Phi} \cdot \left(\frac{s+1}{2}\right) \cdot \frac{\sqrt{q \cdot (1-q)}}{r}\right]^2 \tag{6.158}$$

con Φ-*1(·)* como la inversa de la función de distribución acumulativa normal (= cumulative distribution function, CDF).

Los valores de $I > 5$ y, por tanto, claramente superiores a 1 indican fuertes autocorrelaciones que requieren un examen más detenido. Esto puede deberse a valores iniciales erróneos, parámetros equivocados para el algoritmo MCMC o a un modelo estadístico problemático que genera la distribución apuntada. Los parámetros comunes para q (cuantil), s (probabilidad) y r (= precisión) son q = 0.25, s = 0.95 y r = $\pm$0.005. La salida de `raftery.diag()` da los valores aplicados para q, r y s y, por cadena MCMC, los valores para M (= número de burn-ins sugeridos), N (= tamaño de muestra requerido), N_{min} (= tamaño de muestra mínimo basado en autocorrelaciones 0) e I (= factor de dependencia). De `raftery.diag()`, también se obtiene el número de iteraciones que deben descartarse al principio.

El diagnóstico de Raftery-Lewis difiere en función del cuantil *q* elegido y tiende a ser conservador, es decir, sugiere más iteraciones de las realmente necesarias. Además, sólo examina la convergencia marginal de cada parámetro. Además, este diagnóstico se considera sensible a los pequeños cambios. Pero también tiene la reputación de funcionar bastante bien, especialmente con modelos sencillos (`ptII_quan-_Bayes_HMC.r`).

```
> cbind(chain=rep(1:nchains,each=2), Output
    + do.call("rbind",lapply(raftery.diag(OUTmcmc.list),
    + function(x) x$resmatrix)))
   chain M N     Nmin I
q1 1     8 10084 3746 2.69
q2 1     8 10114 3746 2.70

q1 2     8 11114 3746 2.97
q2 2     8  9630 3746 2.57

q1 3     8 10216 3746 2.73
q2 3     8 10566 3746 2.82

q1 4     8 10834 3746 2.89
q2 4     8 10116 3746 2.70

q1 5     8 10468 3746 2.79
q2 5     8 10780 3746 2.88
```

Basándonos en lo anterior, no hay nada especial de lo que informar aquí – *I* es menor que 5 y *M*, los burn-ins, son más bien pocos.

6.13.3.5 Análisis mediante gráficos de autocorrelación

Los gráficos de autocorrelación muestran para las series temporales hasta qué punto los elementos individuales dependen unos de otros. Una disminución rápida de las dependencias indica que los valores de las series temporales son independientes. Si la curva no se aplana, existen altas autocorrelaciones. En R, esta idea se puede mostrar numéricamente con `autocorr.diag()` o como gráfico con `autocorr.plot()`. La figura 6.110 muestra un gráfico de este tipo para una cadena MCMC (`ptII_quan_Bayes_HMC.r`).

```
> # autocorrelation
> # all chains
> autocorr.diag(OUTmcmc.list)
        q1       q2
Lag 0   1.00000  1.00000
Lag 1   0.53415  0.53462
Lag 5   0.10446  0.10503
Lag 10  0.02425  0.02520
Lag 50 -0.00651 -0.00427
> first MCMC chain
> autocorr.diag(OUTmcmc.list[[1]])
        q1       q2
Lag 0    1.0000  1.0000
Lag 1    0.5428  0.5406
Lag 5    0.1000  0.0981
Lag 10   0.0209  0.0223
Lag 50 -0.0120 -0.0153
> autocorr.plot(OUTmcmc.list[[1]])
```

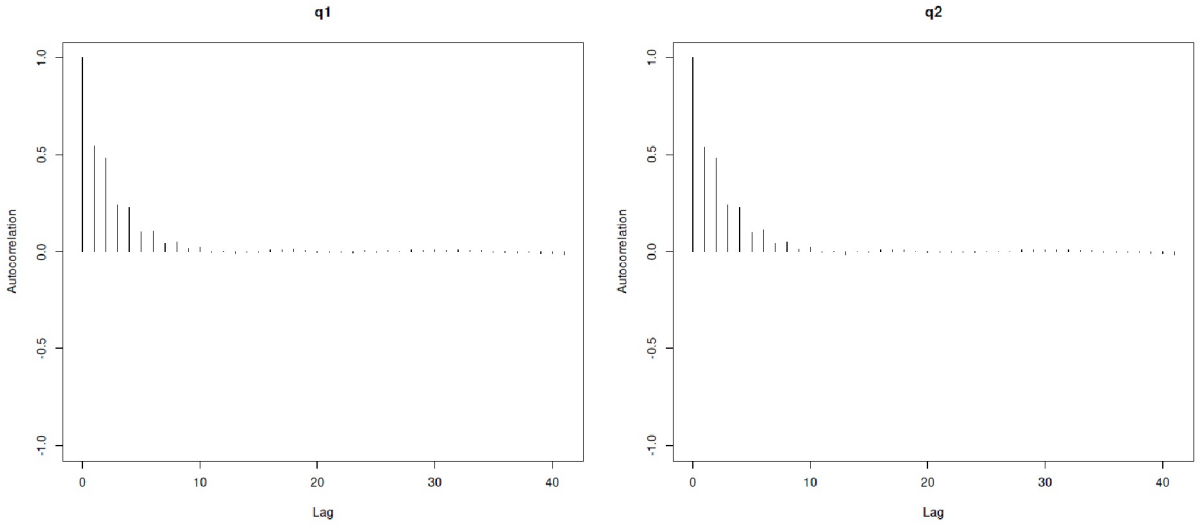


Figura 6.110. *Simulación HMC distribución normal bivariante*
(MCMC de diagnóstico, diagrama de Gelman, ϵ = 0.1, L = 11)

Como puede verse tanto en el gráfico como numéricamente, las autocorrelaciones de los parámetros disminuyen rápidamente al aumentar la distancia, y así es como debería ser. Esto también es cierto para todas las cadenas MCMC cuyos resultados no se imprimen aquí.

6.13.3.6 Tamaño efectivo de la muestra

Como debería quedar claro ahora, las autocorrelaciones son un grave problema en la cadena MCMC, ya que afectan negativamente a la precisión de las estimaciones posteriores como medias, varianzas y cuantiles. El tamaño de muestra efectivo (= effective sample size, ESS) es la estimación del tamaño de muestra necesario para alcanzar la misma precisión que con una muestra aleatoria completa en comparación con las muestras ponderadas o autocorrelacionadas. El método permite resumir la información contenida en los datos empíricos. El método se aplica en simulaciones MCMC de la estadística bayesiana,

series temporales y encuestas. La definición matemática es

$$\mathrm{ESS} = N/D \tag{6.159}$$

donde *N* es el tamaño original de la muestra y *D* es el efecto del diseño (Kish, 1965). El efecto del diseño es especialmente relevante en los estudios de encuestas. Surge como cociente de las varianzas de dos estimadores de un parámetro de interés. Concretamente, por ejemplo, un valor de varianza empírico $s^{2em}{}_{p.}$ se divide contra uno de una muestra (hipotética) completamente aleatoria $\sigma^{2aleatoria}{}_{=\text{efectiva}}$:

$$D = s^2_{\text{emp.}}/\sigma^2_{\text{random}} \tag{6.160}$$

Esto proporciona una medida de la adecuación de una muestra, que es muy relevante en los estudios de encuestas. Para los diagnósticos MCMC, el concepto puede interpretarse como si una cadena MCMC simulada emerge o no como una secuencia de valores aleatorios. Del mismo modo, se compara la varianza de la muestra empírica con la de una muestra aleatoria completa. Kruschke (2015b, p.184) calcula la ESS como el cociente entre el tamaño de la muestra original y la empírica:

$$\mathrm{ESS} = \frac{N}{1 + 2 \cdot \sum_{k=1}^{\infty} \mathrm{ACF}(k)} \tag{6.161}$$

N corresponde al tamaño de la muestra de la cadena MCMC y ACF(k) es la autocorrelación de la cadena MCMC en el intervalo (temporal) k (= lag). Por razones prácticas, la suma se detiene hacia ∞ en cuanto ACF(k) < 0.05, ya que normalmente ACF(k + 1) < ACF(k). Como puede verse, las fórmulas para calcular el ESS no son completamente idénticas ni uniformes. La función R `effectiveSize()` del paquete R `coda` trabaja con la función `spectrum0.ar()`, que utiliza un modelo de series temporales AR (AR = modelo autorregresivo) para estimar la densidad espectral a frecuencia 0. La densidad espectral dividida por la longitud de la serie temporal da la varianza σ^{20} de la media a frecuencia 0. El ESS se estima entonces según

$$\mathrm{ESS} = \frac{N \cdot s^2}{\sigma_0^2} \tag{6.162}$$

donde $s^2 = var(X)$ es la varianza de la muestra y N es el número de repeticiones de la cadena MCMC. Normalmente, las muestras de las cadenas MCMC están correlacionadas positivamente entre sí, es decir, $\sigma^{20} > s^2$. Sin embargo, si las muestras están correlacionadas negativamente entre sí, entonces $\sigma^{20} < s^2$. En este caso, el ESS se hace mayor que N, lo que puede provocar irritaciones en la salida por parte de los usuarios.

En el caso de las simulaciones MCMC, el ESS es por tanto particularmente importante en el caso de autocorrelaciones y falta de convergencia de la cadena MCMC, el ESS es significativamente menor que el tamaño real de la muestra. En otros casos la ESS es una medida de cuánta información independiente contienen las cadenas MCMC autocorrelacionadas. La ESS combina el número de muestras independientes $N_{efectivo\ =\ independiente}$ con las muestras autocorrelacionadas $N_{dependiente}$ – bajo el supuesto de que ambas tienen el mismo poder, es decir, poder explicativo. Por lo tanto, en el caso de las cadenas MCMC autocorrelacionadas, se necesita un tamaño de muestra mucho mayor del necesario en comparación con una cadena MCMC completamente independiente sin ninguna autocorrelación, ya que los solapamientos o redundancias entre los datos reducen el poder explicativo de los datos individuales. Las dos variables son idénticas si no existen autocorrelaciones, lo que es poco probable que ocurra en la práctica. Esto se puede observar en la extracción de una distribución normal con `rnorm()`, ya que el algoritmo de `rnorm()` garantiza que se produzcan extracciones independientes (`ptII_quan_Bayes_HMC.r`):

```
> effectiveSize(rnorm(1e7))
var1
1e+07
```

En este caso, el tamaño de la muestra original corresponde a la EES. Las autocorrelaciones se excluyen completamente. Si se aplica `effectiveSize()` a una lista de cadenas MCMC como un objeto de MCMC de la clase `mcmc.list` esperado por `coda`,

```
> attr(OUTmcmc.list,"class")
[1] "mcmc.list"
```

se suman las ESS individuales.

```
> # effective sample size
> # per chain
> # we have to deal with two parameters, so we are
> # only interested in the number of rows
> OUTmcmc.list.dim <- sapply(OUTmcmc.list, dim)
> OUTmcmc.list.l <- OUTmcmc.list.dim[1,]
> OUTmcmc.list.l
[1] 29918 29918 29918 29918 29918
```

```
> # ESS per chain
> ESS.per.chain <- sapply(OUTmcmc.list, effectiveSize)
> ESS.per.chain
   [,1] [,2] [,3] [,4] [,5]
q1 6297 6095 6461 6234 6187
q2 6399 6206 6286 6258 6289

> # ESS over all chains
> effectiveSize(OUTmcmc.list)
q1    q2
31274 31437

> # ratio compared to actual empirical sample size
> OUTmcmc.list.l/ESS.per.chain
   [,1] [,2] [,3] [,4] [,5]
q1 4.75 4.91 4.63 4.80 4.84
q2 4.68 4.82 4.76 4.78 4.76

> # inverse
> 1/(OUTmcmc.list.l/ESS.per.chain)
    [,1] [,2]  [,3]  [,4]  [,5]
q1 0.210 0.204 0.216 0.208 0.207
q2 0.214 0.207 0.210 0.209 0.210
> # = design effect
> 1/ ( effectiveSize(OUTmcmc.list)/sum(OUTmcmc.list.l) )
q1   q2
4.78 4.76
```

Kruschke (2016b, 2015b, capítulo 7.5.2) recomienda un valor de ESS > 10 000 para obtener una estimación razonable del HDI del 95 %. Otros autores (McElreath, 2015, p.255) asumen estimaciones fiables ya a partir de un valor de ESS de 1000 o menos (por ejemplo, 200), concretamente cuando se trata de estimar medias posteriores. Si existe una distribución posterior normal, sólo se necesitan unos pocos órdenes de magnitud más para una estimación útil de la varianza, mientras que en el caso de una distribución sesgada es necesario considerar qué lado de la distribución es de interés. Sin embargo, si se desea que la estimación posterior sea correcta hasta el percentil 99, el número necesario de ESS vuelve a aumentar drásticamente. En resumen, depende de cuál sea el objetivo de una investigación.

Kruschke (2016b) también señala en una entrada de blog que cuando se trata de combinar valores de ESS de diferentes parámetros de interés, el valor de ESS se reduce debido a las combinaciones de parámetros y es menor que cuando se considera individualmente. Pone el ejemplo de la *d de Cohen*, que se compone de las diferencias medias y las desviaciones típicas agrupadas y, por tanto, resulta de una combinación de parámetros. Si `q1` y `q2` se incluyeran posteriormente en un cálculo, esto no tendría necesariamente consecuencias debido a una correlación existente entre los parámetros, como señala Kruschke. El autor demuestra esta situación utilizando código R en un caso de estudio con las mismas correlaciones de los parámetros (dos valores medios). A pesar de la misma correlación, cada uno de ellos tiene un ESS diferente tras su combinación (diferencia de medias). Esta diferencia se debe al hecho cómo surgieron realmente los valores (valores medios) en cada caso. Por lo tanto, esta información del proceso no se pierde.

Aquí en el ejemplo, `q1` y `q2` se correlacionan con $\varrho = 0.8$, superando el criterio de Kruschke, por lo que la ESS es suficientemente grande. Sin embargo, si se compara esto (véase la salida de R más arriba) con el tamaño real de la muestra, la ESS sigue siendo bastante pequeña. En la combinación de cadenas MCMC, esta circunstancia desempeña entonces un papel menos relevante. En simulaciones informáticas largas con muchas variables, sin embargo, sí desempeña un papel, ya que hay que volver a calcular diferentes modelos en cada caso y el tiempo necesario puede ser considerable.

6.13.3.7 Resumen de los diagnósticos MCMC

En resumen, cabe preguntarse por qué se necesita una prueba estadística clásica que incluya un valor *p* para evaluar y comprobar las cadenas MCMC en un contexto bayesiano, en el que se utilizan preferentemente algoritmos MCMC. Estamos de acuerdo con Kruschke (2012a, 2013c), que ya subraya en el contexto de las comprobaciones predictivas posteriores (véase el capítulo 6.8.4.3) que estas son necesarias, pero deberían ser de naturaleza bayesiana.

- Con respecto a los gráficos y diagnósticos descritos (Gelman-Rubin o Gelman-Rubin-Brooks, Geweke, Heidelberger-Welch y Raftery-Lewis), la edición gráfica del gráfico Gelman-Rubin-Brooks y los diagnósticos Gelman-Rubin parecen especialmente útiles, en parte porque siguen siendo bayesianos.
- El diagnóstico de Geweke como prueba z clásica necesita algunos argumentos para un contexto bayesiano y se podría reconstruir tal vez en una especie de *d* de Cohen: menos pruebas, más atención a una diferencia más o menos estandarizada. Aparte de esto está hay el análisis del error estándar y las interpretaciones sustantivas de ambas variables. Porque el problema básico de las comparaciones estandarizadas – la abstracción y el distanciamiento del contexto del contenido – no debería repetirse.
- Por otra parte, la prueba de Heidelberg-Welch no funciona cerca de 0 en términos de prueba de precisión. Además, es una prueba clásica con un valor *p* y, sin duda, podría reconstruirse bayesianamente incluyendo el intervalo de confianza.
- Puede decirse que Raftery-Lewis es una buena introducción al número de burn-ins, siempre que no se tome la salida al pie de la letra, sino que se observen conscientemente los valores de las cadenas MCMC y se entienda la salida resumida simplemente como una pista.
- El análisis de autocorrelaciones y el tamaño efectivo de la muestra (ESS) son indicadores importantes de la progresión y convergencia de las cadenas MCMC y deberían comprobarse de forma rutinaria sin tener que justificar nunca este paso.

Sin embargo, todas las pruebas, además de otros diagnósticos principalmente gráficos, no son totalmente decisivas, sino que simplemente proporcionan indicaciones en el camino hacia un diagnóstico cuidadoso de las cadenas MCMC con respecto a su adecuación y convergencia. Su conjunto, más todas las demás herramientas de diagnóstico como las comprobaciones predictivas posteriores (s. cap. 6.8.4.3), las interpretaciones del contenido y el recurso a supuestos teóricos, etc. deben considerarse decisivas para determinar si una simulación MCMC con Posterior asociado parece utilizable o no.

6.13.4 Caso práctico: Fisher recargado – más té

En la posguerra de 1946, George Orwell (1946-01-12) escribió un ensayo sobre el té en el Evening Standard titulado "A Nice Cup of Tea". En él comentaba la décima de sus once reglas,

> „Tenthly, one should pour tea into the cup first. This is one of the most controversial points of all; indeed in every family in Britain there are probably two schools of thought on the subject. The milk-first school can bring forward some fairly strong arguments, but I maintain that my own argument is unanswerable. This is that, by putting the tea in first and stirring as one pours, one can exactly regulate the amount of milk whereas one is liable to put in too much milk if one does it the other way round."

La cuestión del orden del té y la leche va mucho más allá del famoso experimento del té de Fisher (véase el capítulo 4.3.2.1). No en vano se puede investigar en términos bayesianos, a los que llegaremos a continuación (`ptII_quan_Bayes_Fisher_LadyBristol-BUGS.r`).

6.13.4.1 Enfoque frecuentista

Recordemos que el análisis de Fisher se realizó mediante una prueba de permutación exacta, ya que Lady Bristol sabía de antemano cuántas veces se servía el té antes que la leche y viceversa. La investigación empírica suele trabajar con condiciones menos restrictivas, es decir, que los sujetos del estudio no conocen las condiciones experimentales tan precisas como en el experimento de Fisher. El conocimiento de Lady Bristol hace que la probabilidad de acertar de ocho tazas para seleccionar correctamente cuatro tazas de la misma variedad utilizando el coeficiente binomial $\binom{8}{4} = 70$ (= número de combinaciones). Esto corresponde a una probabilidad de

```
1/choose(8,4)
```

es decir, $p = 0.0143$, lo que se interpreta clásicamente como *superaleatorio*. Si se conocen cuatro tazas de la misma variedad, las otras cuatro copas resultan automáticamente de la suma marginal fija. Si Lady Bristol no hubiera tenido ninguna información sobre la distribución de las muestras de té de Fisher (técnicamente se desconocen las sumas marginales de la tabla de aptitudes resultante), la prueba binomial exacta `binom.test()` se sugeriría a sí misma. Esto se aplica, por ejemplo, al caso de reconocer las ocho muestras de té o, por ejemplo, seis de ocho muestras cuando se desconoce qué muestra ocurre con qué frecuencia. Sin embargo, dado que Lady Bristol disponía de toda esta información, en principio le bastaba con identificar correctamente las muestras de té hasta disponer de cuatro muestras de una variante. Esto requiere un mínimo de cuatro y un máximo de siete muestras. Las probabilidades binomiales serían por tanto

```
> # frequentist solution to the problem
>
> # min 4 max 7 cups of tea to get 8 of 8 right
> # if there is 4:4 distribution of milk before/after tea
> for(i in 1:8)
+ {
+ if(i == 1)
+ {
+ cat("\nLady Bristol exact binomial probability (one-sided test)\n\n")
+ }
+ cat(i,"of 8 chosen properly with p = ",
+ binom.test(x=i,n=8,p=0.5, alternative="greater")$p.value,"\n")
+ }
Lady Bristol exact binomial probability (one-sided test)
1 of 8 chosen properly with p = 0.9960938
2 of 8 chosen properly with p = 0.9648438
3 of 8 chosen properly with p = 0.8554688
4 of 8 chosen properly with p = 0.6367187
5 of 8 chosen properly with p = 0.3632813
6 of 8 chosen properly with p = 0.1445313
7 of 8 chosen properly with p = 0.03515625
8 of 8 chosen properly with p = 0.00390625
```

y oscilan entre $p = 0.996$ para "adivinar" correctamente una vez y $p = 0.0039$ para identificar correctamente las ocho. Ahora bien, mientras que la mayoría de las fuentes afirman que Lady Bristol reconoció correctamente seis de las ocho muestras (o tres de cuatro de una variante), Salsburg (2001, p.8) proporciona la información

> „And the lady tasting tea, what happened to her? Fisher does not describe the outcome of the experiment that sunny summer afternoon in Cambridge. But Professor Smith told me that the lady identified every single one of the cups correctly“.

Por lo tanto (véase más arriba) existe una probabilidad de $p = 0.0039$ (prueba unilateral) de reconocer las ocho muestras sin error si no se conocen las sumas marginales. Para seis de las ocho muestras, esta probabilidad aumenta a $p = 0.145$, si la hipótesis nula supone $p_{adivinar} = 0.5$ por ejecución. Cambiar a la prueba exacta de permutación de Fisher (también hipótesis direccional, es decir, prueba unilateral) arroja entonces una p = 0.145 para $p = 0.243$ para seis de las ocho correctas. Para los dos casos con una prueba bilateral se observa los resultados de la prueba binomial

```
> binom.test(x=8, n=8, p=0.5, alternative="two.sided")$p.value
[1] 0.0078125
> binom.test(x=6, n=8, p=0.5, alternative="two.sided")$p.value
[1] 0.2890625
```

Para la prueba exacta de Fisher obtenemos

```
> # data table
       > ladybristol.8x8 <- matrix(c(4,0,0,4), nrow=2,
       + dimnames=list(trial=c("milk", "tea"), real=c("milk", "tea")))
> ladybristol.6x8 <- matrix(c(3,1,1,3), nrow=2,
       + dimnames=list(trial=c("milk", "tea"), real=c("milk", "tea")))
> fisher.test(ladybristol.8x8, alternative="greater")$p.value
[1] 0.01428571
> fisher.test(ladybristol.6x8, alternative="greater")$p.value
[1] 0.2428571
```

La hipótesis nula de la prueba exacta de Fisher examina la independencia de las filas de las columnas para tablas 2x2 con el supuesto de odds ratio (= OR) = 1. Mientras que para el caso de seis de ocho muestras correctamente identificadas, no habría que rechazar la hipótesis nula basándose en el frecuen-tismo, pero en el caso de ocho muestras correctas habría que rechazarla con $p = 0.0143$. Por lo tanto, las filas y las columnas dependen unas de otras y la OR es mayor que 1. La OR, calculado mediante la relación de los productos diagonales de la tabla de frecuencias, es para el caso ocho de ocho

```
# odds ratio for 2x2 table of counts (contingency table)
OR.2x2 <- function(tab)
{
nom <- prod(tab[1,1], tab[2,2])
denom <- prod(tab[1,2], tab[2,1])
return(nom/denom)
}
> OR.2x2(ladybristol.8x8)
[1] Inf
```

y para el caso seis de ocho

```
> OR.2x2(ladybristol.6x8)
[1] 9
```

Hasta aquí la situación frecuentista.

6.13.4.2 El enfoque bayesiano

En el caso de Bayes, se añade el conocimiento previo y se plantea la cuestión de cómo se configura el proceso de selección binomial o de toma de decisiones (= parámetro del modelo de Likelihood) para Lady

Bristol si imaginamos que hacemos más experimentos con ella. La cuestión aquí es qué capacidades tiene realmente Lady Bristol. El supuesto $p_1 = 0.5$, por ejemplo, representa la probabilidad del modelo binomial. Todo lo que sea menos que eso requiere una explicación de por qué alguien acertaría peor que el azar. Un evaluador externo ficticio, tras observar a Lady Bristol y sus habilidades, podría llegar a la conclusión de que Lady Bristol puede adivinar con una probabilidad de $p_2 = 0.8$ e identificar correctamente las muestras de té de acuerdo con el criterio "*la leche antes del té*" o no. Tanto las conjeturas como las observaciones del evaluador externo representan la Likelihood, es decir, el proceso generador de datos, que aquí corresponde a un proceso de Bernoulli.

Esto complementa el conocimiento previo. Debe entenderse independientemente de la capacidad real de Lady Bristol, aunque psicológicamente existe un solapamiento que aquí ignoramos deliberadamente. Por ejemplo, la habilidad y la creencia en ella en uno mismo están muy relacionadas, como pueden demostrar muchos estudios y casos prácticos. Sólo que esto se manifiesta de forma diferente, dependiendo de la confianza fundamental en uno mismo, de la forma del día, de un posible resfriado que dificulte las papilas gustativas, del nerviosismo, etc. Desde el punto de vista del Fisher más escéptico, podemos suponer un conocimiento previo de $p_{Fisher} = 0.2$ de que Lady Bristol tiene esta habilidad. Aquí no nos preocupa la probabilidad de acertar del 50% por carrera, sino una evaluación a priori de las habilidades de Lady Bristol.

Fisher obviamente no confía en Lady Bristol para hacer esto. Lady Bristol, por otro lado, puede tener tener un nivel muy alto de confianza en sus propias habilidades; después de todo, bebe mucho té todos los días y ha adquirido bastante experiencia a lo largo de los años con la cuestión de "*la leche antes del té o no*". Tal vez sea simplemente natural. Su conocimiento previo, es decir, la expectativa del resultado del experimento, se estima en $p_{Bristol} = 0.9$, que sigue siendo significativamente superior a las hipótesis del evaluador externo sobre el proceso de probabilidad. El escéptico Fisher sabe poco sobre Lady Bristol y sus insospechadas habilidades con el té. En este sentido que él, como científico, expresa una expectativa previa mucho menor sobre el resultado del experimento. Esto demuestra que el conocimiento previo no surge simplemente de un capricho subjetivo sino que se simplifica como resultado de una cuidadosa deliberación, y siempre se basa en una información o conocimiento del observador. Sólo el empirismo, es decir datos reales, conduce a la larga a una convergencia de estos puntos de vista inicialmente tan diferentes. O bien Fisher aprende que Lady Bristol es más capaz de lo que él esperaba; o bien Lady Bristol aprende que sus capacidades no son tan amplias como ella esperaba – o bien ambos aprenden lo mismo al mismo tiempo. En realidad, probablemente fue Fisher quien aprendió algo sobre Lady Bristol y Lady Bristol se encontró con la confirmación, pero no aprendió nada. Su nivel de conocimiento, a diferencia del de Fisher, se mantuvo más bien constante.

En lo que sigue, seguimos a Stefan y Schönbrodt (2017). La tabla 6.11 resume la situación. Es importante distinguir la columna $p_{individual}$ y la columna p_{BT} La primera columna proporciona las opiniones individuales previas de Fisher y de Lady Bristol. Para el teorema de Bayes y el denominador, la probabilidad total, estos dos valores deben deben combinarse. Aquí cubren todo el espacio de hipótesis y se contradicen. Es decir, Fisher es Fisher y no Lady Bristol y viceversa. Sin embargo, ambos sostienen puntos de vista diferentes, que junto con 0.2 + 0.9 = 1.1 superan el valor de 1, que marca el límite natural de las probabilidades, ya que la probabilidad total no puede superar 1. Así que las dos opiniones previas de Fisher y Lady Bristol se escalan a su suma con respecto a 1. Este valor a priori p_{BT} se utiliza entonces en el teorema de Bayes. Este procedimiento permite que cada persona estime intuitivamente su conocimiento previo en 1 como el límite superior de una probabilidad; y, sin embargo, los valores pueden entrar en el teorema de Bayes independientemente como valores relativos.

Tabla 6.11: *Lady Bristol's Teeskills*

	Likelihood	Proceso de Bernoulli (generador de datos)	p	
	$\text{Likelihood}_{\text{adivinar}}$	adivinar	0.5	
	$\text{Likelihood}_{\text{Experto}}$	Experto externo (convencido por Lady Bristol)	0.8	
Caso	Prior	Expectativas a priori de destrezas de Lady Bristol	$p_{\text{individal}}$	p_{BT}
1	Prior_{Fisher}	Fisher (muy escéptico)	0.1	0.1432
1	$\text{Prior}_{Bristol}$	Experto (más convencido de Lady)	0.6	0.857
Caso	Prior	Expectativas a priori de destrezas de Lady Bristol	$p_{\text{individual}}$	p_{BT}
2	Prior_{Fisher}	Fisher (skeptisch)	0.2	0.182
2	$\text{Prior}_{Bristol}$	Marido de Lady Bristol (convencido por ella)	0.9	0.818

No lo necesitamos para la Likelihood. Ésta representa el proceso de generación de datos y no tiene por qué sumar 1 en términos probabilísticos e independientemente de la conjetura o del evaluador externo. Ambos describen procesos diferentes, que se examinan por separado (`ptII_quan_Bayes_Fisher_-LadyBristo-BUGS.r`).

En primer lugar, especificamos el conocimiento previo. Distinguimos dos casos: un Fisher escéptico con una Lady Muriel Bristol confiada y un Fisher menos escéptico con una Lady Muriel muy confiada. A continuación, consideramos el caso 2, el Fisher menos escéptico y la Lady Muriel ciertamente muy motivada.

```
#example
0.2/(0.2+0.9) #=0.182 Fisher
0.9/(0.2+0.9) #=0.818 Lady Bristol
#
#
# case 1 R-Code
# Fisher is very skeptical that Lady Bristol
# can do it with p = 0.5 (!p = 1-0.5 = 0.5)
# he thinks she is guessing
# Muriel herself is rather confident that she
# can do it with p = 0.8 (!p = 1-0.8 = 0.2)
# not run
#P.Fisher.hypo <- 0.5
#P.Muriel.hypo <- 0.8
# case 2
# Fisher is less skeptical that Lady Bristol
# can do it with p = 0.6 (!p = 1-0.6 = 0.4)
# Muriel is highly confident that she
# can do it with p = 0.9 (!p = 1-0.9 = 0.1)
P.Fisher.hypo <- 0.6
P.Muriel.hypo <- 0.9
> # IF AND ONLY IF THERE ARE NO OTHER INFOS WE CAN DO THE FOLLOWING
> # re-scale prior knowledge to sum up to p = 1
> P.Fisher.priorH1 <- P.Fisher.hypo/(P.Fisher.hypo + P.Muriel.hypo)
> P.Muriel.priorH2 <- P.Muriel.hypo/(P.Fisher.hypo + P.Muriel.hypo)
```

El punto de vista a priori apunta siempre a los supuestos de Fisher y Lady Muriel respectivamente.

```
> P.Fisher.priorH1
[1] 0.4
> P.Muriel.priorH2
[1] 0.6
```

Alternativamente, podríamos asignar probabilidades directamente, cosa que no hacemos:

```
# OTHERWISE WE CAN ADD PRIOR PROBS TO EACH HYPOTHESIS
# IMPORTANT - THEY HAVE TO SUM UP TO p = 1
# case we assume both hypos are both equally probable
#P.Fisher.priorH1 <- 0.5
#P.Muriel.priorH2 <- 0.5
# case we assume Fisher is more realistic than Lady Bristol
#P.Fisher.priorH1 <- 0.7
#P.Muriel.priorH2 <- 0.3
# case we are confident that Lady Bristol knows what she talks about
# and Fisher is a non-believer
#P.Fisher.priorH1 <- 0.1
#P.Muriel.priorH2 <- 0.9
```

A partir de ahí, podemos calcular un factor de Bayes previo, una odds ratio.

```
> # prior Odds Ratio, some kind of Bayes Factor = BF_prior
> BF.prior.skeptic <- P.Fisher.priorH1/P.Muriel.priorH2
> BF.prior.skeptic
[1] 0.6666667
> 1/BF.prior.skeptic
[1] 1.5
```

Ahora sigue el proceso de generación de datos, la probabilidad binomial (= Likelihood). Partimos con los datos empíricos y tomamos los valores reales de la realidad (Salsburg, 2001), no los de la literatura establecida.

```
# empirical data
# case 1
# arbitrary values from literature
#successes <- 2
#ntrials <- 8
# case 2
# values from literature
#successes <- 6
#ntrials <- 8
# case 3
# true values
successes <- 8
ntrials <- 8
# failures
failures <- ntrials - successes
```

y procedemos al cálculo de la probabilidad para el primer caso – ocho aciertos de ocho intentos:

```
> # likelihood under prior hypos =
> # expectations = prior knowledge = whatever...
> # likeli for binomial = (n over k) * p^s * q^f =
> # (n over k) * p^s * (1-p)^(n-s)
> # (n over k) = constant
> const <- choose(ntrials,successes)
> # =
> # const <- exp(lchoose(ntrials,successes))
> # =
> # const <- exp(lfactorial(ntrials) -
> # lfactorial(successes) - lfactorial(failures))
> const
```

```
[1] 1
>
> L.Fisher <- const * P.Fisher.priorH1^successes *
+ (1-P.Fisher.priorH1)^(failures)
> L.Muriel <- const * P.Muriel.priorH2^successes *
+ (1-P.Muriel.priorH2)^(failures)
> L.Fisher
[1] 0.00065536
> L.Muriel
[1] 0.01679616
```

El denominador del teorema de Bayes – la probabilidad o evidencia total – se obtiene a partir de la suma de las dos hipótesis previas H_1 y H_2 multiplicada por las Likelihoods de las hipótesis respectivas.

```
> totalprob.case1 <- P.Fisher.priorH1*L.Fisher +
+ P.Muriel.priorH2*L.Muriel
> totalprob.case1
[1] 0.01033984
```

De este modo, podemos comparar las dos visiones de las capacidades de Lady Muriel, la probabilidad de adivinación frente a la del evaluador externo. Las Likelihoods bajo las respectivas hipótesis sobre las capacidades de Lady Muriel dan como resultado otro factor de Bayes (cociente de Likelihood).

```
> # LR = likelihoodratio, another kind of Bayes Factor = BF
> LR.FM <- L.Fisher/L.Muriel
> LR.MF <- L.Muriel/L.Fisher
> LR.FM
[1] 0.03901844
> LR.MF
[1] 25.62891
```

Ahora sigue una actualización del conocimiento previo hacia la Posterior multiplicando los factores de Bayes previos por la proporción de Likelihood. El resultado son dos factores de Bayes para cada una de las dos hipótesis respecto a las habilidades de Lady Muriel, cada uno de los cuales refleja los cambios en las hipótesis previas de Fisher (= baja confianza en Lady Muriel) y Lady Muriel (= alta confianza en sí misma). Los cambios se refieren al cociente de las dos Likelihoods (ninguna capacidad, alta capacidad).

```
> # posterior probs according to Bayes Theorem
> post.case1.FisherH1 <- P.Fisher.priorH1*L.Fisher / totalprob.case1
> post.case1.MurielH2 <- P.Muriel.priorH2*L.Muriel / totalprob.case1
> post.case1.FisherH1
[1] 0.02535281
> post.case1.MurielH2
[1] 0.9746472
```

Los resultados no son sorprendentes. Si se hace una suposición a priori más alta, esto tiene un efecto en la Posterior y viceversa. Una hipótesis a priori más baja dará lugar a una hipótesis a posteriori más baja. Las Posterior-Odds son entonces

```
> # posterior Odds
> post.case1.FisherH1 / post.case1.MurielH2
[1] 0.02601229
> post.case1.MurielH2 / post.case1.FisherH1
[1] 38.44336
```

y la comparación de los dos factores de Bayes para las hipótesis previas originales

```
> # update posterior odds ratio by prior believe
> # and likelihood ie. data
> # LR combined with prior knowledge beyond a flat prior
> BF.posterior.hypoF.vs.hypoM.S <- P.Fisher.priorH1/
+ P.Muriel.priorH2 * LR.FM
> BF.posterior.hypoF.vs.hypoM.S
[1] 0.02601229
> # =
> BF.prior.skeptic*LR.FM
[1] 0.02601229
> 1/BF.posterior.hypoF.vs.hypoM.S
[1] 38.44336
```

Como era de esperar, la opinión previa de Lady Muriel tiene una probabilidad relativa significativa-mente mayor que la de Fisher en el momento posterior. Con el aumento de las pruebas, es decir, de los datos empíricos, estas opiniones convergerán cada vez más qua actualización.

Otra variante (Rheem, 2017) también examina cómo cambia la expectativa inicial ante los datos empíricos. Esta vez tiene lugar inicialmente sin factores de Bayes y sólo con el teorema de Bayes. El proceso generador de datos sigue siendo el mismo proceso binomial. Sin embargo, esta vez fijamos la Likelihood en las capacidades de Lady Bristol (parámetro del modelo binomial) con $p_{no\ especificado} = 0.5$, dada la falta de información. La hipótesis prior H_1 determina $p_{H1} = 0.5$, es decir, que Lady Bristol sólo puede adivinar, mientras que la hipótesis prior alternativa H_2 especifica $p_{H2} = 0.7$, pero no totalmente de su capacidades. Hemos combinado los pasos individuales anteriores en la función de R `fishertest.BT()` (`ptII_quan_Bayes_Fisher_LadyBristol-BUGS.r`).

```
> ladymuriel.BT(pr1=0.5, pr2=0.7, si=8, Ni=8)
Lady Muriel via Bayes Theorem
prior1 = 0.417
prior2 = 0.583
BF.pr12 = 0.714
BF.pr21 = 1.4
const = 1
successes = 8
failures = 0
trials = 8
L1 = 0.000908
L2 = 0.0134
evidence = 0.0082
LR12 = 0.0678
LR21 = 14.8
post1 = 0.0462
post2 = 0.954
postOR12 = 0.0484
postOR21 = 20.7
NOTE: Prior values adjusted to sum up to 1
```

Como puede observarse, las probabilidades posteriores de ambas hipótesis cambian de prior a posteriores. Las dos hipótesis a priori pueden estar directamente relacionadas entre sí como factores de Bayes (Prior Odds). Si cambiamos los valores empíricos a 6 en lugar de 8 éxitos, obtenemos

```
Lady Muriel via Bayes Theorem R-Output
prior1 = 0.417
prior2 = 0.583
BF.pr12 = 0.714
BF.pr21 = 1.4
const = 28
successes = 6
failures = 2
```

```
trials = 8
L1 = 0.0499
L2 = 0.192
evidence = 0.132
LR12 = 0.26
LR21 = 3.84
post1 = 0.157
post2 = 0.843
postOR12 = 0.186
postOR21 = 5.38
NOTE: Prior values adjusted to sum up to 1
```

Con las mismas expectativas a priori, resultados empíricos diferentes conducen – lógicamente – a valores posteriores diferentes. Si utilizamos los valores posteriores como nuevas expectativas previas y nos quedamos con 6 aciertos, los valores posteriores y las Posterior-Odds ya cambian:

```
> res.ft.BT <- ladymuriel.BT(pr1=0.5, pr2=0.7, si=6, Ni=8, prout=FALSE)
> ladymuriel.BT(pr1=res.ft.BT[["post1"]],
+ pr2=res.ft.BT[["post2"]],
+ si=6, Ni=8)
Lady Muriel via Bayes Theorem
prior1 = 0.157
prior2 = 0.843
BF.pr12 = 0.186
BF.pr21 = 5.38
const = 28
successes = 6
failures = 2
trials = 8
L1 = 0.000296
L2 = 0.247
evidence = 0.209
LR12 = 0.0012
LR21 = 837
post1 = 0.000222
post2 = 1
postOR12 = 0.000222
postOR21 = 4500
```

Si repetimos esto unas cuantas veces, sólo podremos calcular en la escala logarítmica, porque los valores ya son demasiado pequeños (o demasiado grandes). Sería una tarea para los lectores reescribir la función R `ladymuriel.BT()` en la escala logarítmica.

Si especificamos una Prior uniforme plana a1 = 1 y b1 = 1 de acuerdo con lo que hemos dicho hasta ahora, la Posterior es Beta(a1 + s, b1 + n - s) = Beta(1 + 8, 1 + 8 - 8) = Beta(9, 1). En ese caso, podríamos plantearnos si una Prior uniforme representa correctamente nuestro estado de conocimiento. No sabemos si Lady Muriel miente, o no tiene ni idea sobre el té, o es demasiado confiada. Pero en cada prueba tiene una probabilidad de acertar del 50%, por lo que la probabilidad global de respuestas correctas tiende a disminuir con cada intento. Dado que en la prueba exacta de Fisher, el número de ensayos y el número de resultados correctos/incorrectos se conocen de antemano, la Prior debe ajustarse en consecuencia – si suponemos que la probabilidad de acertar a priori es del 50 %. Por lo tanto, una Prior uniforme no es una buena opción (véase Studer, 1996b), por eso elegimos una Haldane-Prior, es decir, una especie de distribución pre-priori.

```
> # calculation via beta update
> theta <- seq(0,1,0.01)
> # data 8 of 8
> si <- 8
> Ni <- 8
> # Haldane prior
> a.prior1 <- 0.5
> b.prior1 <- 0.5
```

```
> # beta update
> ab.post1 <- bino.ab.post(a.prior=a.prior1, b.prior=b.prior1,
        + si=si, Ni=Ni)
> ab.post1
a b
8.5 0.5
attr(,"type")
[1] "post"
> beta.dens1 <- dbeta(theta, shape1=ab.post1["a"],
        + shape2=ab.post1["b"])
```

resumimos la Posterior

```
> # summaries
> postbeta1 <- beta.summary(a=ab.post1["a"], b=ab.post1["b"])
> t(unlist(postbeta1))
     a    b    mode    mean      sd         var
[1,] 8.5 0.5 1.071429 0.9444444 0.07243558 0.005246914
```

y trazarla (véase la Fig. 6.111 arriba) como

```
# plot
plot.siNi(theta, beta.dens=beta.dens1, pbl.dens=muriel.pbl,
        pjc.dens=muriel.pjc, a.prior=a.prior1, b.prior=b.prior1)
```

Para otro estado de conocimiento a priori, buscamos los parámetros a priori para a_1 y a_2. Elegimos una distribución beta que tiene su moda en $\theta = 0.5$. Con `beta.determine.opt()` esto es sencillo

```
> # determine prior beforehand
> betadet.ab2 <- beta.determine.opt(p=c(0.5,0.99999,0.00001),
+ qua=c(0.5,0.95,0.05), ab.start=NULL, graph=TRUE)
> # betadet.ab2
> a.prior2 <- betadet.ab2[["res.ab"]]["a"]
> b.prior2 <- betadet.ab2[["res.ab"]]["b"]
> a.prior2
a
5.687128
> b.prior2
b
5.687128
```

y obtenemos $a_2 = 5.687$ y $b_2 = 687$ para los valores a priori de la distribución beta. Según las reglas anteriores, esto da como resultado una Beta(5.687 + 8- 5.687 + 8 - 8) = Beta(13.687, 5.687). La figura 6.112 muestra la distribución beta a priori que estamos buscando. Expresada en R más el gráfico de la Posterior (véase la Fig. 6.111 abajo)

```
# beta update
ab.post2 <- bino.ab.post(a.prior=a.prior2, b.prior=b.prior2,
        si=si, Ni=Ni)
beta.dens2 <- dbeta(theta, shape1=ab.post2["a"],
        shape2=ab.post2["b"])
# plot
plot.siNi(theta, beta.dens=beta.dens2,
        pbl.dens=muriel.pbl, pjc.dens=muriel.pjc)
```

y el resumen de la Posterior:

```
> # summaries
> ab.post2
a b
13.687128 5.687128
attr(,"type")
[1] "post"
> postbeta2 <- beta.summary(a=ab.post2["a"], b=ab.post2["b"])
> t(unlist(postbeta2))
     a        b        mode      mean      sd        var
[1,] 13.68713 5.687128 0.7302257 0.7064595 0.1008874 0.01017826
```

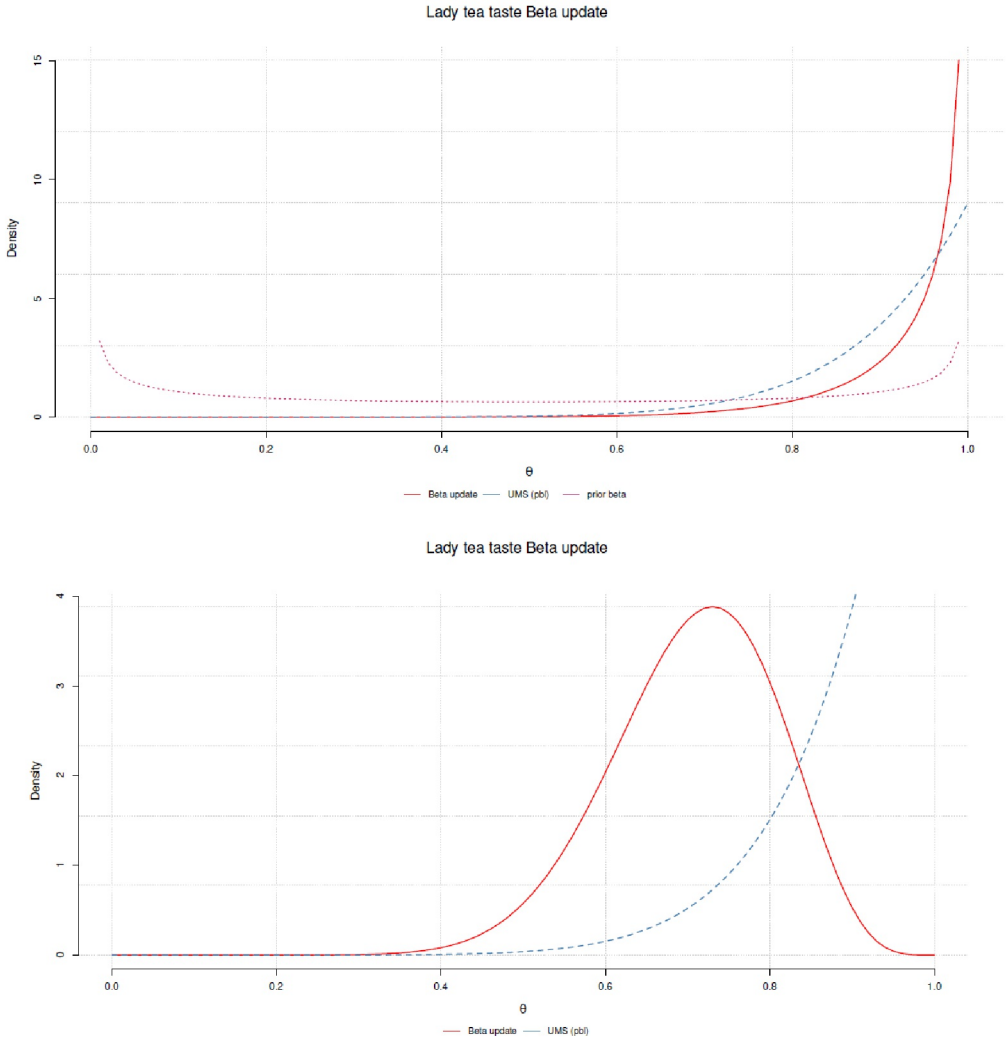


Figura 6.111. *Lady Muriel (Beta Update, Haldane Prior, 8 de 8 éxitos)*

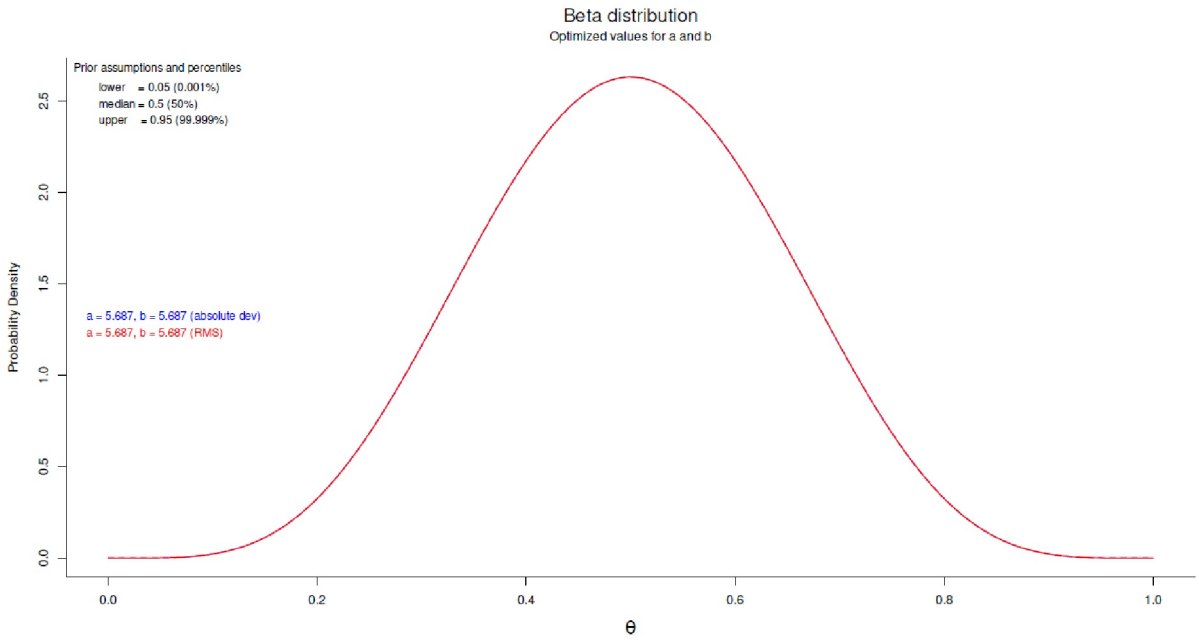


Figura 6.112. *Lady Muriel (Beta Plot)*

La odds ratio de la Posterior (Zellner & Siow, 1980; Hicks, Rodríguez-Campos & Choi, 2017), que *no debe confundirse* con el factor de Bayes, da como resultado

```
> p1 <- postbeta1[["mean"]]
> p2 <- postbeta2[["mean"]]
> # simple ratio
> p1/p2
[1] 1.33687
> 1-(p1/p2)
[1] -0.3368698
> # post OR
> (p1 * (1-p2)) / (p2 * (1-p1))
[1] 7.063657
```

6.13.4.3 Lady Muriel y MCMC con BUGS

Aunque una solución analítica es fácil de realizar aquí, intentaremos una aproximación numérica de las integrales mediante simulación MCMC. Lunn, Jackson, Best, Thomas y Spiegelhalter (2012, p.122) resuelven el problema de Lady Bristol mediante BUGS/ MCMC sin, no obstante, mantener constantes ambos márgenes de la tabla (ibíd., p.126). Esto tiene como consecuencia que se determina el problema como prueba binominal. Hemos escrito una función BUGS (véase más abajo) que, basado en la distribución hipergeométrica, mantiene constantes los márgenes de la tabla y, por lo tanto, corresponde a la prueba exacta de Fisher.

```
# based on hypergeometric distribution
model8 <- c("
model {
for (i in 1:2) {
```

```
y[i] ~ dhyper(n[1], m1, N, psi)
}
psi <- (p[1] * (1-p[2])) / (p[2] * (1-p[1]))
p[2] ~ dunif(0, 1)
p[1] ~ dunif(0, 1)
N <- n[1] + n[2]
m1 <- y[1] + y[2]
y[1] ~ dbin(p[1], n[1])
y[2] ~ dbin(p[2], n[2])
post <- step(p[1] - p[2])
}
")
```

Como proceso binomial generador de datos, suponemos, al igual que los autores (ibíd.), $p = 0.5$ y suponemos (lo erróneo) 6 de 8 ensayos con éxito para Lady Bristol (`ptII_quan_Bayes_Fisher_-LadyBristol-BUGS.r`). Dejamos los demás modelos a un lado, los autores los enumeran con salida. Así que elegimos el 50% como Prior.

```
# Lady Bristol gets six out of eight cups right
dats6of8 <- c("
list(n=c(4,4), y=c(3,1))
Initial values for independent probabilities
(though gen.inits is sufficient)
list(p=c(0.5, 0.5))
")
```

En primer lugar, adoptamos una postura neutral con un prior uniforme.3 Para la secuencia de llamadas, seguimos a Neto (2014). El siguiente código R se ejecutó bajo Windows 7™ porque hubo problemas con la compilación del paquete R `BRugs` bajo Debian Bullseye.

```
> # run.model() taken and adopted from
> # http://www.di.fc.ul.pt/~jpn/r/bugs/bugs_tutorial.html
> # it brings BUGS calls together
> # run model 8 with 6 of 8
> run.model(model=model8, samples=c("post","p","psi"), dats=dats6of8)
model is syntactically correct
data loaded
model compiled
initial values generated, model initialized
10000 updates took 0 s
monitor set for variable 'post'
monitor set for variable 'p'
monitor set for variable 'psi'
100000 updates took 0 s
> samplesStats("*")
     mean    sd    MC_error val2.5pc median  val97.5pc start sample
p[1] 0.667   0.179 0.000316 0.28360  0.6861  0.9471    10001 300000
p[2] 0.333   0.178 0.000334 0.05286  0.3134  0.7166    10001 300000
post 0.897   0.304 0.000569 0.00000  1.0000  1.0000    10001 300000
psi 16.100 178.400 0.328200 0.42410  5.0220 90.2400    10001 300000
```

El modelo con 8 éxitos proporcionó:

```
> # run model 8 with 8 of 8
> run.model(model=model8, samples=c("post","p","psi"), dats=dats8of8)
model is syntactically correct
data loaded
model compiled
initial values generated, model initialized
10000 updates took 0 s
```

```
monitor set for variable 'post'
monitor set for variable 'p'
monitor set for variable 'psi'
100000 updates took 0 s
> samplesStats("*")
       mean  sd       MC_error val2.5pc  median val97.5pc start sample
p[1]   0.834 1.41e-01 2.65e-04 0.477300  0.8708    0.9950 10001 300000
p[2]   0.167 1.41e-01 2.55e-04 0.005001  0.1293    0.5216 10001 300000
post   0.996 6.33e-02 1.24e-04 1.000000  1.0000    1.0000 10001 300000
psi 2564.000 2.83e+05 5.14e+02 2.529000 53.3800 4992.0000 10001 300000
```

A partir del modelo BUGS, se pueden extraer las cadenas MCMC para los parámetros individuales

```
# use coda from here on...
bugs.out.mcmc <- buildMCMC("*")
str(bugs.out.mcmc)
```

y analizarlos posteriormente, por ejemplo con el paquete R `coda`, `BEST` o `bayesplot`. Las posibilidades de análisis de las simulaciones MCMC también están disponibles en `BRugs`, a saber, utilizando el ejemplo del parámetro `post`. Los parámetros marcados con * se aplican a todos los parámetros del modelo:

```
# single plots for single parameters
plotBgr("post")
plotAutoC("post")
plotBgr("post")
plotDensity("post")
plotHistory("post")
bugs.out <- samplesHistory("*", plot=FALSE)
str(bugs.out)
mean(bugs.out$post)
```

Lo que funciona en BUGS también funciona en JAGS. En algunos casos, sin embargo, la sintaxis o la llamada a la función se realiza de forma diferente. Un ejemplo es la función `dhyper()` para calcular la distribución hipergeométrica no central – necesaria para la prueba con sumas marginales fijas (Plummer, 2017-06-28, p.50 o p.69). Aquí prescindimos de ello.

Un caso especial en la elección de distribuciones a priori es el principio de máxima entropía, que se remonta a los trabajos de Jaynes (1957a, 1957b) y Claude Shannon sobre la teoría de la información (1948). A continuación nos ocuparemos de ello.

6.14 Entropía máxima

La entropía máxima tiene su origen en los campos de la termodinámica y la mecánica estadística (Jaynes, 1963, 1988b). El concepto ha tenido importantes influencias en la teoría de la información, la estadística y el aprendizaje automático. La entropía máxima en física denota el estado de un sistema en el que prevalece el mayor desorden (= entropía) y, por tanto, no se evidencia una orientación clara ni un centro de gravedad de la materia contenida en él y su energía asociada, que está más allá de la información conocida. La entropía máxima denota, por tanto, la ausencia de estructuras con contenido de información, es decir, el *estado actual de error*.

En la termodinámica (2ª ley), la entropía denota el objetivo de un sistema sin más aporte de energía. La entropía se denota con términos como desorden, aleatoriedad e incertidumbre. Desde el punto de vista de la teoría de la información, la entropía máxima representa el modelo estadístico que mejor representa el

estado actual del conocimiento en comparación con otros modelos que contienen la misma información. Por tanto, contiene una combinación de información conocida (conocimiento previo) e incertidumbre máxima (con respecto a todos los demás factores de entrada desconocidos).

Cualquiera que sea la información adicional desconocida, la entropía se maximiza y se asume un estado de ignorancia. Esto explica por qué la entropía máxima es una opción legítima y bien fundamentada en el *caso de la máxima información incierta* para la elección de una Prior; y lo mismo ocurre con las distribuciones a priori informadas. Se incluye en el modelo la información que se conoce y todo lo demás se ajusta al estado de máxima entropía para no introducir sesgos en el modelo. Esto proporciona un estado de máximo aprendizaje a partir de los datos empíricos.

Jaynes (1988b) distingue el teorema de Bayes del principio de máxima entropía de la siguiente manera:

- La aplicación del teorema de Bayes calcula la probabilidad posterior de una hipótesis *H* dados los datos *D* y la información previa disponible *I* además de los datos, es decir $p(H|DI) = p(H|I) * p(DjHI)/p(D|I)$. Esto suele implicar estimaciones de parámetros, ya que *H* puede ser una afirmación sobre una propiedad de un parámetro θ.
- La aplicación del principio de máxima entropía asigna una distribución $(p_1; \ldots ; p_n)$ a un espacio de hipótesis $(H_1;\ldots; H_n)$ con el criterio de maximizar la entropía de la información, es decir $S_i = -\Sigma p_i * log(p_i)$. Esto confiere a la distribución la propiedad deseada de maximizar la entropía. Así, si se ha utilizado toda la información disponible, toda la ambigüedad restante se resuelve, es decir, se modela, mediante la maximización de la entropía. La entropía máxima es, por tanto, un modelo de *cómo* debería ser una distribución según un determinado criterio – la entropía máxima – después de que toda la información disponible haya fluido hacia ella. La distribución resultante corresponde a una distribución de probabilidad sobre el espacio de hipótesis *H*. Según el principio, la distribución prácticamente se forma sola, de modo que no se requiere ninguna otra entrada numérica aparte de la información que fluye hacia ella de antemano.

6.14.1 ¿Qué es la información?

Las características de la información no siempre son intuitivas. Basándonos en la definición de Bateson (1985) de que *la información es la diferencia que hace la diferencia*, es decir, que *conduce al significado*, se pueden derivar criterios adicionales más formales para la máxima entropía y la estadística bayesiana. Se pueden derivar criterios. Si la información *I* de un suceso *i* ocurre con probabilidad p_i se cumple que (Keng, 2017)

- $I(p_i)$ – La información aumenta con la probabilidad decreciente de un suceso y viceversa.
- $I(p_i) \geq 0$ – Información se denota sólo en el rango positivo. No hay información negativa.
- $I(p_i = 0) = 0$ o $I(p_i = 1) = 0$ – Los eventos con probabilidades de exactamente 0 y 1 son triviales y no llevan información. No están definidos.
- $I(1 \geq p_i \geq 0) \geq 0$ – Los sucesos con información justo por encima de 0 sí conllevan información sustancial – cuando realmente se produce un (único) suceso. Por tanto, la sorpresa de que ocurra algo inesperado conlleva información. Pero incluso si no se produce un suceso casi seguro justo por debajo de uno, esto genera una experiencia de aprendizaje que contiene información.
- Lo ideal es que la información se distribuya simétricamente y en contra de la probabilidad de un suceso. Con una probabilidad media del 50%, el rendimiento informativo es muy alto (ejemplo: moneda justa) y la entropía (véase más abajo) llega al máximo. Cuantas más distorsiones entren en juego, menor es el rendimiento informativo, puesto que el resultado de un suceso ya no depende del número máximo de estados y entradas y las probabilidades se mueven hacia un suceso cada vez más cierto o cada vez más incierto. Cuanto más cierto o incierto es un suceso, menos inputs contribuyen a su ocurrencia y menos información es necesaria para explicar la ocurrencia de un suceso.
- $I(p_i, p_j) = I(p_i) + I(p_j)$ – Los sucesos independientes proporcionan información independiente, que a su vez puede sumarse. Es decir, la recopilación de información de tales sucesos independientes da lugar a la misma cantidad de información, independientemente de que se recojan uno tras otro o todos a la vez.

Con estas premisas, podemos pasar al concepto de entropía *H*, que se puede definir un poco más adelante para distribuciones de probabilidad discretas y continuas.

6.14.2 ¿Qué es la entropía?

¿Qué significa la entropía en nuestra realidad? Muy sencillo: mire la habitación de un niño después de haberla ordenado completamente. Entonces se puede ver el desarrollo de la entropía casi en tiempo real. Del mismo modo, un juguete del tipo que se agita delante de los gatos es adecuado para que te persigan. Esperas unos segundos de uso y luego cuentas los enredos, nudos y otras cosas que ya se han creado durante este tiempo. A partir de ahí el tiempo necesario para ordenar de nuevo la habitación de los niños o para desenredar los juguetes del gato, desenredar todos los nudos, etc. De este modo, se obtiene rápidamente una impresión realista de lo que es realmente la entropía.

Desde el punto de vista de la *termodinámica*, la entropía describe de forma bastante aburrida el estado de un sistema, de manera que cuando el sistema absorbe energía (calor), la entropía aumenta y cuando libera energía (calor), disminuye. La entropía también puede aumentar a través de procesos internos espontáneos del sistema, es decir, a través de todos los procesos que implican una conversión de energía en energía térmica. Si un sistema está cerrado y aislado del entorno, la entropía no puede disminuir sino que, según la segunda ley de la termodinámica, sólo puede aumentar con el tiempo. En la *mecánica estadística*, la entropía es el número de microestados energéticamente equivalentes y disponibles en un macroestado dado. Una mayor entropía de los macroestados significa más microestados y éstos son entonces estadísticamente más probables que los macroestados con menor entropía. Si un sistema no se modifica desde el exterior, es decir, se deja a abandonado a sí mismo, la entropía a largo plazo se aproxima al macroestado que tiene la entropía más alta a la misma energía en comparación con todos los demás macroestados. Así, el estado de máxima entropía es el objetivo final de muchos microestados diferentes, siempre que el sistema se deje a sí mismo. La entropía es, por tanto, una medida de la falta de conocimiento sobre el estado de los elementos individuales del sistema, o expresada en términos de tecnología de la información de tal manera que apenas es posible extraer conclusiones de un estado de máxima entropía a nivel macro al microestado del mismo sistema. Además, en vista de la entropía máxima, no es posible hacer afirmaciones claras sobre el camino recorrido hacia la entropía máxima. La entropía es una medida de la ignorancia. La entropía se vuelve particularmente impresionante en el contexto de un agujero negro, que por definición pierde toda posibilidad de reconstrucción de la información posible con el cruce del horizonte de sucesos, y por tanto existe una incertidumbre máxima-total respecto al estado en el interior de un agujero negro. La entropía de Bekenstein-Hawking define la cantidad de entropía para una descripción termodinámicamente adecuada de un agujero negro (mejor su superficie calculable, el área del horizonte de sucesos) por un observador externo. Esto conduce, mediante desvíos, reducciones cuánticas, antipartículas, etc., a la radiación de Hawking y a la suposición de que, sin más aporte de energía, un agujero negro emite radiación (los pequeños bastante más que los grandes) y, por tanto, se encoge y, si no es infinito, se disuelve en un tiempo "previsible". Este último es un periodo de tiempo previsible ciertamente largo, pero no por ello menos calculable. Un agujero negro del tamaño de nuestro sol tendría una duración de 1064 años, que está mucho más allá de cualquier observación. Un agujero negro con una temperatura inferior a la radiación cósmica de fondo no se encoge porque las condiciones termodinámicas no hacen que el agujero negro emita "calor" en forma de radiación a su entorno. Esto sólo ocurre cuando las condiciones se invierten, es decir, cuando la temperatura ambiente del agujero negro es inferior a la del agujero negro. De ello se deduce que el estado informativo-teórico de los distintos agujeros negros es diferente, pero no determinable.

En la mecánica estadística, Ludwig Boltzmann (1844-1906) dedujo de ello la *fórmula de entropía de Boltzmann*.

$$S = k_B \cdot \ln(\Omega) \qquad (6.163)$$

S es la entropía de un sistema cerrado. La cantidad Ω representa el número de microestados dentro de este sistema, descritos mediante conceptos compatibles con la termodinámica (por ejemplo, la ubicación y el momento de todas las partículas de un sistema). Entre ellos se incluyen la energía, el volumen y el número de partículas. La cantidad Ω combina estos conceptos. El factor k_B representa la constante de Boltzmann, una constante natural introducida por Max Planck (1858-1947) en honor a Boltzmann. Al igual que la entropía, tiene la unidad Joule por Kelvin, es decir, energía por unidad de temperatura. La fórmula de la entropía de Boltzmann puede extenderse a los sistemas. Si p_i son las probabilidades de los microestados *i*, la entropía *S* del macroestado es

$$S = -k_B \cdot \sum_i p_i \cdot \ln(p_i) \tag{6.164}$$

También se puede imaginar que, sin influencia externa, la probabilidad de que en un sistema se formen espontáneamente estructuras con contenido informativo es significativamente menor que la probabilidad de que esto no ocurra, ya que el conjunto de estados no estructurales es mucho mayor que el conjunto de estados estructurales con claro contenido informativo. Desde el punto de vista entrópico, saber algo sobre el macroestado no significa saber al mismo tiempo algo sobre el microestado (véase el excursus sobre los agujeros negros más arriba). La cosa se complica por el hecho de que tanto un sistema termodinámico como el estado de entropía requieren un conjunto exacto de parámetros ajustados para poder ser descritos exhaustivamente en cada caso. Si aquí se siente de repente el reproche de la subjetividad – a saber: ¿no existe una descripción y una relación objetivas de los sistemas y los parámetros? -- la respuesta es que vivimos en un mundo de verdad relativa. Se trata de describir lo que es relativo en el sentido de causa-efecto y no lo que es absoluto. Lo que se desprende de esta supuesta subjetividad es que la Entropía Máxima es un enfoque metodológicamente controlado y gobernado por reglas para establecer una Prior para el teorema de Bayes después de haber vertido en él *toda* la información contextual disponible.

Según la teoría de la información de Claude Elwood Shannon (1916-2001), la entropía *H* se define desde aproximadamente 1948 como el valor esperado del contenido de información *I(p)*, si *p* = probabilidad de aparición de un elemento del sistema con contenido informativo principal.

$$H = \sum_{k \epsilon K} p_z \cdot I(p_k) \tag{6.165}$$

Expresada en una curva, la entropía Boltzmanniana y Shannoniana se ven como en la Fig. 6.113 (arriba, abajo):

```
# information content = surprisal
# https://en.wikipedia.org/wiki/Entropy_(information_theory)
# that's identical
# -log2 p(event) = log(1/ p(event))
ps <- seq(0,1,.01)
par(oma=c(2,1,2,1), "cex.axis"=1, bty="l", mfrow=c(2,2))
plot(ps,-log2(ps),col="violetred3", type="l",
       pre.plot=grid(), bg="darkred",
       main=expression(paste("-",log[2],"(p)",sep="")),
       bty="n", xlab="p(E)",
       ylab=expression(paste("-",log[2],"(p(E))"),sep=""))
plot(ps,log(1/ps),col="violetred3", type="l",
       pre.plot=grid(), bg="darkred",
       main=expression(paste("log(1/p)",sep="")),
       bty="n", xlab="p(E)",
ylab=expression(paste(log,"(1/p(E))"),sep=""))
ps1 <- seq(0,3,0.01)
plot(ps1,ps1*log(ps1), col="violetred3", type="l",
pre.plot=grid(), bg="darkred",
       main=expression(paste("p * log(p)",sep="")),
       bty="n", xlab="p",
       ylab=expression(paste("p * log(p)"),sep=""))
```

```
ps1 <- seq(0,3,0.01)
plot(ps1,ps1*log2(ps1), col="violetred3", type="l",
pre.plot=grid(), bg="darkred",
       main=expression(paste("p * ",log[2],"(p)",sep="")),
       bty="n", xlab="p",
       ylab=expression(paste("p * ",log[2],"(p)"),sep=""))
mtext("Information content of an event E with p(E)",
       outer=TRUE, line=-0.5, cex=1.5, side=3)
```

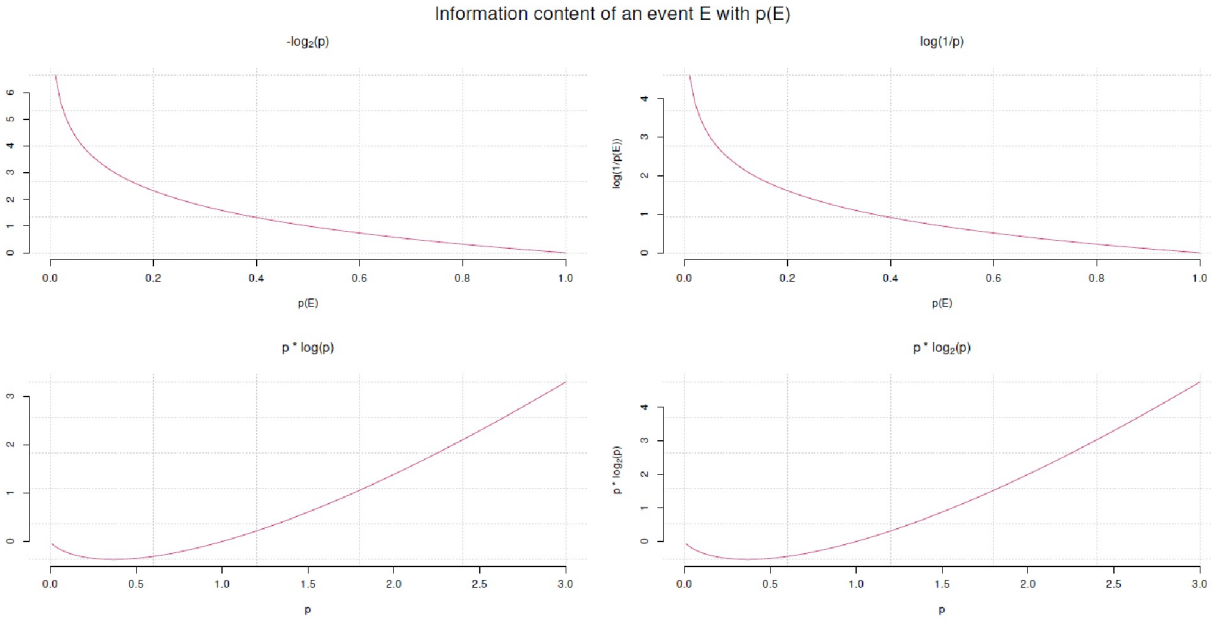


Figura 6.113. *Máxima entropía (contenido informativo de un suceso)*

Los elementos del sistema con baja probabilidad de aparición no contribuyen prácticamente nada a la información general. Entropía en el sentido de desorden e información son precisamente las perspectivas 180 grados desplazadas del mismo fenómeno. Un alto contenido de información requiere que desorden sea mínimo y una entropía máxima requiere la presencia de información mínima, es decir, máxima incertidumbre. Como ejemplo, basta con observar la curva de entropía de una moneda justa en la Figura 6.114, una curva de forma U invertida. La máxima entropía termina en un valor de $p = 0.5$ y $q = 1 - p = 0.5$. En este punto, hay un lugar en el que existe la máxima incertidumbre sobre lo que ocurrirá a continuación. El contenido informativo de la única de muchos lanzamientos es máximamente incierto o mínimamente informativo sobre lo que ocurrirá concretamente a continuación. La distribución de probabilidad en este caso sigue una distribución uniforme (ptII_quan_Bayes_-MaximumEntropy.r).

```
# entropy of a coin with p = 1-q ie. p != q
p <- 0.5
#-(p*log2(p)+(1-p)*log2(1-p))
H <- function(p) -p*log2(p) -(1-p)*log2(1-p)
# entropy H(X) of a perfect fair coin
H1 <- function(e=NA, base=2)
{
#p = prob of event e - here binary, ie. e=2
p <- 1/e
H <- -sum(replicate(e,p*log(p, base=base)))
return(H)
}
```

Si se utiliza la escala $\log_2$ (= bits), dos o tres sucesos posibles dan como resultado una entropía de

```
> # binary event Shannon entropy
> # two events, same prob for each event to occur
> outcomes <- 2
> # on the bit scale
> log2(outcomes)
[1] 1
>
> # three events, same prob for each event to occur
> outcomes <- 3
> # on the bit scale
> log2(outcomes)
[1] 1.584963
> # i.e. more possible events, the higher the entropy
```

o con la función

```
> # fair coin
> p <- 0.5
> H(p=p)
[1] 1
>
> # entropy of a pure fair coin with p1 = p2 = 0.5
> H1(e=2,base=2)
[1] 1
> H1(e=3,base=2)
[1] 1.584963
```

Y como gráfico (Fig. 6.114 arriba)

```
# entropy of a pure fair coin with p1 = p2 = 0.5
ps <- seq(0,1,0.01)
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l")
plot(ps, H(ps), col="violetred3", type="l", pre.plot=grid(), bg="darkred",
       main="", bty="n", xlab="p(E)",
       ylab=expression(paste("H = p(E) * log(p(E))"),sep=""))
mtext("Shannon entropy of a perfect fair coin", outer=TRUE,
       line=-2, cex=1.5, side=3)
points(p,Hp <- H(p=0.5), cex=1.5, col="darkred",bg="green", pch=21)
text(x=p, y=Hp*0.97, col="black",
       labels=paste("p = ",p," | H(p) = ",round(Hp,dig),sep=""),
adj=1, offset=0.5, pos=1)
```

Si las probabilidades fueran, por ejemplo, $p = 0.3$ y $q = 1- p = 0.7$, el contenido informativo aumentaría considerablemente y la incertidumbre disminuiría. Aparte de eso, esto sería una indicación de que la moneda no es justa. La entropía cambia entonces como muestra la Fig. 6.114 abajo) para varios de estos casos (código R no impreso).

```
> # p not 0.5 but 0.7
> p <- 0.7
> qu <- 1-p
> qu
[1] 0.3
> H(p=p)
[1] 0.8812909
> H(p=p) < H1(e=2,base=2)
[1] TRUE
```

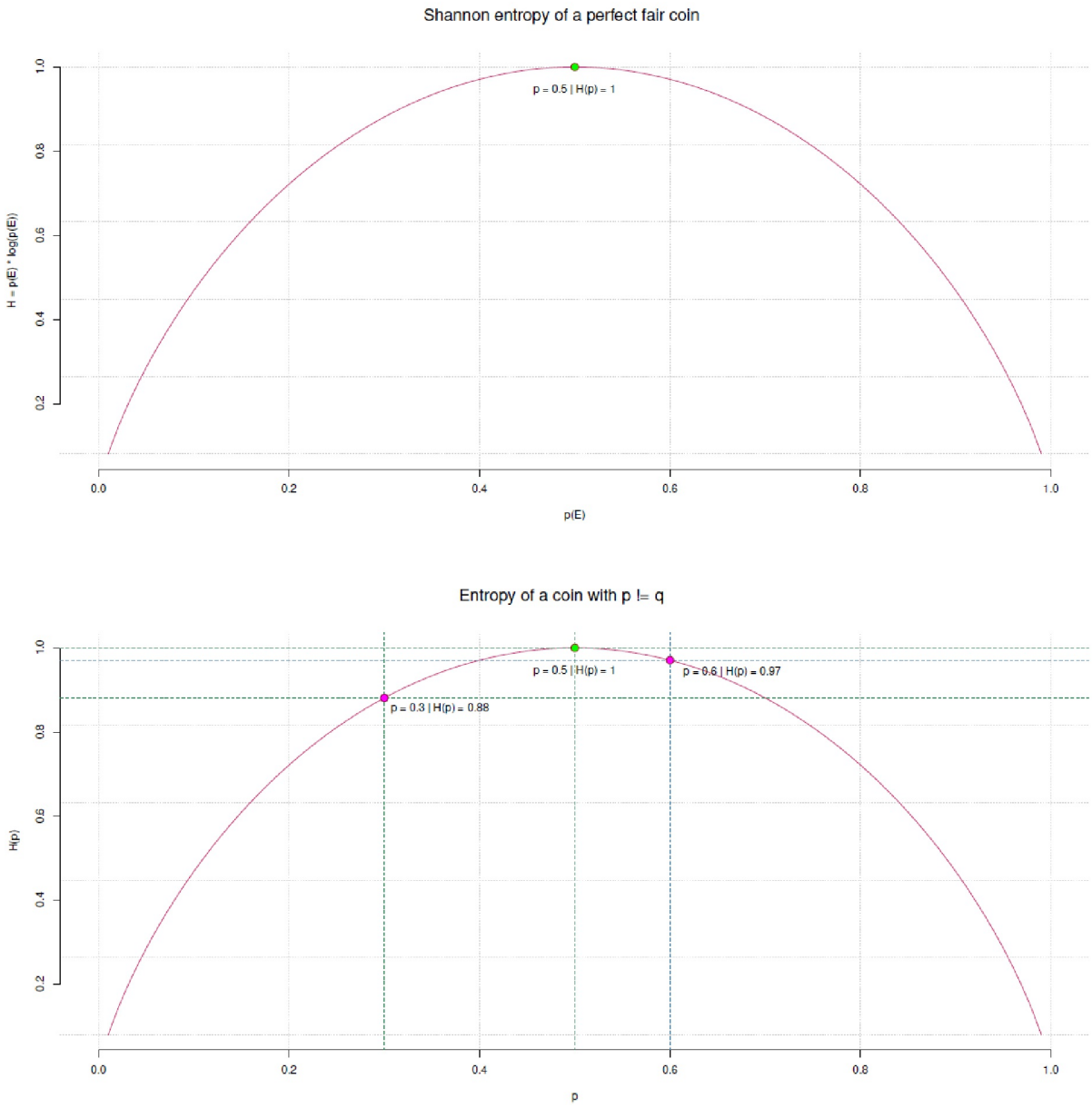


Figura 6.114. *Entropía máxima (entropía de una moneda justa o injusta)*

Por tanto, la entropía (incertidumbre) ha descendido a $H = 0.881$. Si p desciende a $p = 1$ y $q = 0$, es decir, se produce un suceso de certeza máxima, la entropía desciende a 0; y el contenido de información es de certeza máxima con $I = 1$. La figura 6.114 (parte abajo) muestra estos cambios con las líneas discontinuas. Esto corresponde a un caso que no se da en la práctica.

En general, en el caso discreto para la variable aleatoria X y la distribución $Pr(X = x_k) = p_k$, la entropía H se define como

$$\mathrm{H}(X) = -\sum_{k\geq 1} p_k \cdot log(p_k) \qquad (6.166)$$

Para el caso continuo, se aplica lo siguiente a la variable aleatoria X con densidad $p(x)$

$$\mathrm{H}(X) = -\int_{-\infty}^{\infty} p(x) \cdot log(p(x))\,\mathrm{d}x \tag{6.167}$$

Así pues, la entropía crece con el número de posibilidades si esto no va acompañado de un aumento de las estructuras con contenido informativo (véase la Fig. 6.115).

```
# growing possibilities
n <- 1:10
log2(choose(6,k))
Hn <- log2(n)
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l")
plot(n, Hn, col="violetred3", type="b", pch=21, pre.plot=grid(),
       bg="darkred", main="", bty="n",
       xlab="number n of possibilities", ylab="H= log2(n)",
       ylim=c(0,4.5), xlim=c(0,10), axes=FALSE)
axis(2, cex=0.8)
axis(1, c(0,n), labels=as.character(c(0,n)), cex=0.8)
text(2,log2(2),labels="coin",pos=4, cex=0.9, col="darkred")
text(6,log2(6),labels="dice",pos=1, cex=0.9, col="darkred")
mtext("Entropy of growing possibilities",
       outer=TRUE, line=-2, cex=1.5, side=3)
```

Una entrada en Wikipedia (2019i) enumera las distintas distribuciones de probabilidad y sus equivalentes de máxima entropía, así como sus peculiaridades. Llama la atención que la distribución uniforme es la única que es constante bajo Máxima Entropía y no cambia. La distribución uniforme está relacionada con el principio de indiferencia („principle of indifference“) de Laplace y Bernoulli, más tarde conocido como principio de razón insuficiente („principle of insuffcient reason“, Studer, 1996b) - la forma más simple de una Prior no informado (Jaynes, 2003). La figura 6.115 contiene la comparación de esas posibilidades y su entropía debido al aumento del número de posibilidades. Si el número *n* de posibilidades es $n = 2$ se trata de lanzar una moneda al aire (`ptII_quan_Bayes_MaximumEntropy.r`).

```
> log2(2)
[1] 1
```

Si n = 6, se trata de un cubo (=dice):

```
> log2(6)
[1] 2.584963
```

Para cada posibilidad existe una probabilidad individual $p = 1/n$, para la serie *k* resulta p^k . La relación entre el dado y la moneda es

```
> log2(6)/log2(2)
[1] 2.584963
```

es decir, un lanzamiento de dados corresponde a 2.585 lanzamientos de una moneda.

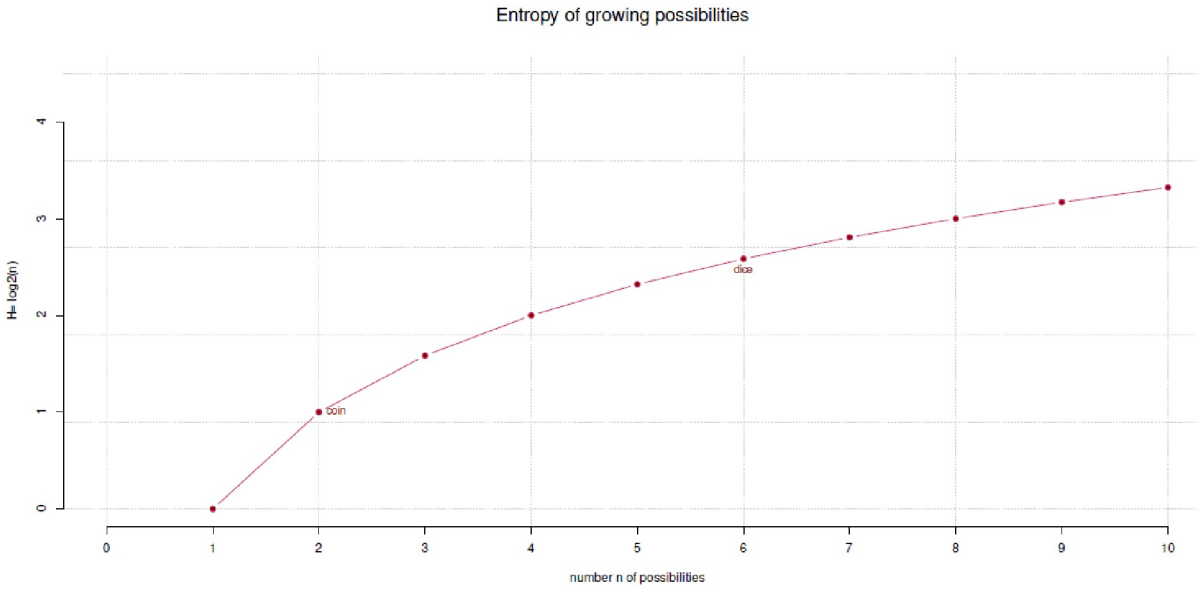


Figura 6.115. *Entropía máxima* (*entropía y crecimiento de las posibilidades)*

La entropía puede utilizarse fuera de la termodinámica y la teoría pura de la información, por ejemplo en el análisis de textos, que a su vez tiene naturalmente cierta proximidad con la teoría de la información.

El análisis de textos trata de la codificación del lenguaje y de la cuestión de cuánta información se comunica en los enunciados lingüísticos. La comparación de entropías permite la clasificación del lenguaje y de los textos, por ejemplo, si una clasificación conduce a una ganancia de información. A continuación se presenta un ejemplo empírico.

6.14.3 Estudio de caso: "Yo, nosotros y la nación" – autoanuncio presidencial

En la ciencia política existe la hipótesis (Sayre, 2001) de que en la campaña electoral presidencial (ejemplo: Estados Unidos de América), el presidente en funciones hace referencia a sus éxitos ("yo") se refiere a sus éxitos ("yo") y el aspirante intenta reunir a la nación ("nosotros", "nación") detrás de él. De esto se deduce que el presidente tiende a utilizar la forma de primera persona orientada al ego y el contrincante tiende a hablar en la forma "nosotros" y debe utilizar términos como "nación". Los debates presidenciales públicos registrados en EE.UU. desde 1988 podrían servir para ilustrar muy bien esta hipótesis (Commission on Presidential Debates, 1987). En el caso de dos aspirantes, por ejemplo porque el presidente no puede ejercer otro mandato, esta diferencia debería desaparecer.

Para simplificar en gran medida el modelo de lenguaje complejo subyacente, los términos (aislados) "yo", "nosotros" y "nación" pueden contarse para cada orador. Para los tres duelos de discursos entre el (entonces) actual presidente George W. Bush y su contrincante John Kerry en 2004, realizamos este trabajo con AQUAD 7 (Gürtler & Huber, 2013; Huber & Gürtler, 2012) (véase el cuadro 6.12). En el capítulo 6.15.3 se pueden examinar los datos más detenidamente. Aquí nos centramos en dos aspectos:

- *Diferenciación significativa* entre Bush vs. Kerry: en relación con las frecuencias de los tres términos, ¿merece la pena la diferenciación entre Bush y Kerry desde el punto de vista de la entropía? Entonces, la diferenciación según locutor conlleva una reducción significativa de la entropía y una ganancia de información?

- *Comparación de entropía* Bush vs. Kerry: Si comparamos las capacidades de los tres terminos en Bush y Kerry, nos preguntamos si, desde el punto de vista de la entropía, existe una diferencia significativa entre Bush y Kerry con respecto al uso de los tres términos.

Para aclarar la primera pregunta, utilizamos la fórmula de entropía de Shannon anterior para calcular la entropía para todo el conjunto de datos, por separado para Bush y Kerry. La entropía se suma según la parte relativa respectiva del total y la diferencia con la entropía total. Este valor de diferencia denota el aumento y la información o la reducción de entropía al introducir la distinción de Bush frente a Kerry.

Para aclarar la segunda cuestión, recurrimos a la *divergencia de Kullback-Leibler* para frecuencias (Kullback & Leibler, 1951). Corresponde a la comparación de dos distribuciones de probabilidad; por lo general, datos empíricos frente a expectativas teóricas. Aquí se comparan datos empíricos entre sí. Es importante comprender el punto de referencia, es decir, desde qué punto de vista se realiza la comparación. Las probabilidades bin de Shannon se sustituyen por relativas empíricas. La fórmula para el cálculo discreto de la divergencia de Kullback-Leibler cuando P y Q son funciones de probabilidad es

$$D_{\mathrm{KL}}(P \,||\, Q) = \sum_{x \epsilon X} P(x) \cdot log\left(\frac{P(x)}{Q(x)}\right) \tag{6.168}$$

$$= -\sum_{x \epsilon X} P(x) \cdot log\left(\frac{Q(x)}{P(x)}\right) \tag{6.169}$$

y para el caso continuo

$$D_{\mathrm{KL}}(P \,||\, Q) = \int_{-\infty}^{\infty} P(x) \cdot log\left(\frac{P(x)}{Q(x)}\right) \mathrm{d}x \tag{6.170}$$

Tabla 6.12: *Bush vs. Kerry (uso de palabras clave)*

	Bush	*Kerry*	$\sum$
nation	16	32	48
I	101	131	232
we	91	88	179
$\sum$	208	251	459

Empecemos por la cuestión de si tiene sentido diferenciar entre Bush y Kerry. La función `H.counts()` calcula la entropía para un vector con frecuencias según la fórmula de Shannon y para una variante bayesiana con una Prior sobre la base de la distribución de Dirichlet (Hauser & Strimmer, 2009; paquete R `entropy`). En primer lugar, se calcula la entropía para Bush, Kerry y combinada para ambos por separado (`ptII_quan_Bayes_Entropy_KullbackLeibler.r`).

```
# presidential debates and entropy
pres <- t(matrix(c(16,101,91, 32,131,88), nrow=2, byrow=TRUE,
dimnames=list(c("Bush","Kerry"),c("nation","I","we"))))
# table used terms for all
bush <- pres[,"Bush"]
kerry <- pres[,"Kerry"]
counts.all <- apply(pres,1,sum)
```

Un vistazo a los datos brutos:

```
> pres.dat
  counts.bk president term
1 16        Bush      nation
2 101       Bush      I
3 91        Bush      we
4 32        Kerry     nation
5 131       Kerry     I
6 88        Kerry     we
> counts.all
nation I   we
48     232 179
```

Ahora pasamos a obtener información distinguiendo entre Bush y Kerry:

```
# Shannon entropy
# https://www.youtube.com/watch?v=IPkRVpXtbdY
Hc.all <- H.counts(counts=counts.all)#pres.dat[,"counts.bk"])
Hc.bush <- H.counts(counts=bush)
Hc.kerry <- H.counts(counts=kerry)
# combined
ratio <- apply(pres,2,sum)/sum(pres)
Hc.comb <- c(sum(ratio * c(Hc.bush["Hc"], Hc.kerry["Hc"])),
sum(ratio * c(Hc.bush["Hc.Bayes"],
Hc.kerry["Hc.Bayes"]))
)
names(Hc.comb) <- c("Hc.comb","Hc.comb.Bayes")
# information gain by distinction (partition)
# of Bush vs. Kerry
info.gain <- Hc.all - Hc.comb
```

A continuación, examinamos los resultados intermedios de la entropía:

```
> ratio
Bush     Kerry
0.453159 0.546841
> info.gain
Hc          Hc.Bayes
0.005820272 0.004540140
> Hc.comb
Hc.comb   Hc.comb.Bayes
0.9423960 0.9447993
```

Los resultados indican que la ganancia de información debida a la distinción entre Bush y Kerry resulta ser muy modesta, es decir, $H\Delta = 0.0058$ y $H_{Bayes}\Delta = 0.0045$ respectivamente para la variante bayesiana con prior estándar "Jeffreys". Esto significa que los dos términos examinados son muy muy similares en cuanto a sus frecuencias. Sin embargo, esto *no dice nada* sobre *cómo* Bush y Kerry utilizan lingüísticamente los términos. Comprender esto sería tarea de una reconstrucción cualitativa. Si añadimos la comparación directa de entropía según Kullback-Leibler a la misma,

```
> KL.counts(bush,kerry)
KL         CHI^2
0.02297473 0.04412671
```

entonces la diferencia de entropía $H(KL)_{B\text{-}vs\text{-}K} = 0.023$ no muestra una gran diferencia. En general, Bush parece ser algo menos "entrópico" que Kerry, lo que confirman los cálculos individuales. Pero, de nuevo, esto no dice nada sobre la forma real de los enunciados lingüísticos del discurso. Es un elemento informativo entre muchos otros. No hay pruebas desde esta perspectiva de que Bush y Kerry difieran significativamente.

```
> Hc.all
Hc        Hc.Bayes
0.9482163 0.9493394
> Hc.bush
Hc        Hc.Bayes
0.9097650 0.9130463
> Hc.kerry
Hc        Hc.Bayes
0.9694369 0.9711125
```

El paquete R `entropy` ofrece funciones comparables y ampliadas para la entropía:

```
> apply(pres,2,entropy)
Bush      Kerry
0.9097650 0.9694369
> entropy(counts.all)
[1] 0.9482163
```

En la sección 6.15.3, ampliamos el ejemplo para explorar esta cuestión de investigación con más detalle utilizando diferentes implementaciones MCMC, la prueba t bayesiana y los factores de Bayes en R, y comparamos los resultados.

6.14.4 Máxima entropía y estadística bayesiana

La máxima entropía se utiliza, como se explicó al principio, cuando prácticamente no se dispone de información contextual o cuando ésta ya se ha incorporado. La entropía máxima sustituye al *principio de razón insuficiente* formulado por Laplace. Un statu quo de estados igualmente probables, que puede representarse mediante una distribución uniforme, recibe una base bien fundada. La distribución uniforme puede describirse técnicamente como un caso especial de la distribución beta *Beta(1, 1)*. En el contexto de los análisis bayesianos, la entropía máxima es principalmente un método para *modelar una distribución a priori*, aunque su ámbito de aplicación es mucho más amplio (Jaynes, 1988b; Gregory, 2006; McElreath, 2015).

La idea rectora es *maximizar la entropía – es decir, el desorden – de la distribución a priori* cuando no se dispone de información sobre la ya utilizada y disponible (contextual) para orientar la Prior (por ejemplo, restricciones, ponderaciones, etc.). Esto se debe a que toda la información disponible y utilizable restringe una distribución a priori, por ejemplo, a una determinada gama de valores y expectativas con respecto a la localización de cuantiles, valor modal, valor medio, etc. Con la máxima entropía, se produce la menor especificación posible con la condición de que todo se haga *por adelantado* para permitir que la información fluya hacia la Prior. Las aplicaciones más comunes y sobre todo físicas son, por ejemplo, la extracción de señales del ruido mejorando drásticamente la relación señal-ruido o el análisis de imágenes para afinar imágenes ruidosas o borrosas (Gregory, 2006, cap. 8 o p.208 para ejemplos de reconstrucción de imágenes ruidosas).

Según las fórmulas anteriores, el Principio de Máxima Entropía se aplica tanto a cantidades discretas como continuas. Más concretamente, se trata de toda una clase de distribuciones de máxima entropía, ya que prácticamente todas, excepto la distribución uniforme, conocen un equivalente de máxima entropía en el que se maximiza la entropía (Wikipedia, 2019i).

Según Jaynes (1988b), el Principio de Máxima Entropía debe utilizarse en la fase exploratoria de la generación de hipótesis, mientras que el Teorema de Bayes conduce después a un análisis numérico del vínculo entre el conocimiento previo y los datos, la Posterior. En este sentido, en lo que respecta a la elección de una Prior, se dispone de un arsenal muy amplio y potente para las distribuciones más diversas con el fin de generar una Prior adecuada.

McElreath (2015) o Jaynes (1957a, 1957b, 1988b) muestran ejemplos de cómo se puede implementar el principio de máxima entropía sobre datos concretos. Studer (1996b) utiliza un ejemplo de tasas de éxito en la terapia clínica de la adicción para mostrar cómo se puede derivar matemáticamente y en términos de contenido una distribución a priori para calcular las tasas de éxito sobre la base de la distribución binomial a lo largo de la máxima entropía. Encontrará una versión resumida del planteamiento de Studer más adelante en el estudio de caso sobre las tasas de aprobados en la terapia de la adicción (véase el capítulo 6.15.2).

6.14.5 El clásico: ¿es justo un dado?

En 1957, E.T. Jaynes (1957a, 1957b, 1963, 1978) llevó a cabo una línea de pensamiento hasta entonces completamente nueva al combinar, en primer lugar, la entropía desde el punto de vista de la termodinámica con la teoría de la información según Claude Shannon y, en segundo lugar, al utilizar el enfoque de la entropía máxima para resolver problemas concretos. Utilizó la teoría de la información para formular una distribución de probabilidad basada en información incompleta y, al mismo tiempo, para formular restricciones, es decir, información previa. Es importante darse cuenta de que todo lo que no sea distribución uniforme equivale a la introducción de restricciones.

A esta formación de conocimiento previo le siguió la aplicación de la máxima entropía para seleccionar entre todas las posibles distribuciones de probabilidad la que cumplía estas restricciones, la que maximizaba la entropía y representaba así el mejor estado de conocimiento. De este modo se garantiza que no se añada ninguna otra información de forma distorsionadora. Luego Jaynes transfirió estos pensamientos a su contexto, la mecánica estadística. El ejemplo clásico de Jayne (1963) preguntaba si un dado es justo si el valor esperado B de los ojos es, por ejemplo, $B = 4.5$. Un dado justo con seis caras tiene una probabilidad por cara de 1/6 cada una y un valor esperado de B = 3.5 (`ptII_quan_Bayes_MaximumEntropy_-Jaynes-fair-dice.r`).

```
> # expectation fair dice (die)
> k <- 1:6
> prob <- 1/6
> sum(k*prob)
[1] 3.5
```

Jaynes parte de la conocida fórmula de entropía de Shannon. A continuación, introduce dos restricciones que representan el estado del conocimiento previo:

1. Las probabilidades suman $p = 1$
2. El valor esperado B es $B = 4.5$.

Estas dos restricciones pueden combinarse con la fórmula de entropía mediante los *multiplicadores de Lagrange* (Studer, 1996b; Peneld, 2003; ambos con ejemplos) y resolverse. El método de los multiplicadores de Lagrange, en honor al matemático y astrónomo italiano Joseph-Louis Lagrange (1736-1813), es un método matemático para resolver un problema de optimización con restricciones. El resultado es un sistema de ecuaciones que puede aproximarse mediante un algoritmo de optimización. Las probabilidades p_i indican aquellas de las caras del cubo, el factor k denota el número de caras del cubo, B representa el valor esperado del cubo, λ_0 y λ_1 representan los multiplicadores de Lagrange de las dos restricciones.

$$\mathcal{L}(p_1, ..., p_6, \lambda_0, \lambda_1) = -\sum p_k \cdot log(p_k) - \lambda_0 \cdot \left(\sum_{k=1}^{6} p_k - 1\right) - \lambda_1 \cdot \left(\sum_{k=1}^{6} k \cdot p_k - B\right) \tag{6.171}$$

Seguimos las sencillas explicaciones y derivaciones de Keng (2017). Se pude implementar las fórmulas fácilmente en R. Para la optimización y determinación de λ se utiliza `optim()`. También sería posible utilizar `optimize()` o `uniroot()`. Para ilustrar cómo se va a implementar el procedimiento a lo largo de las fórmulas, sigue una *aproximación de fuerza bruta*, que prueba y traza obstinadamente los posibles valores y los traza para obtener una aproximación al punto 0 correspondiente (`ptII_quan_Bayes_-MaximumEntropy_Jaynes-fair-dice.r`).

```
# manual optimization
sek <- seq(-3,3,.001)
prob.zeros <- NA
for(i in 1:length(sek))
{
prob.zeros[i] <- foo3(x=sek[i], B=4.5, k=1:6)
}#
find value nearest zero
naid.sm.zero <- which(prob.zeros <= 0)
naid.gr.zero <- which(prob.zeros >= 0)
sm.zero <- prob.zeros[naid.sm.zero][1]
prob.zeros.temp <- prob.zeros[naid.gr.zero]
gr.zero <- prob.zeros.temp[length(prob.zeros.temp)]
MIN <- min(abs(c(sm.zero, gr.zero)))
MIN.id <- sek[which(abs(prob.zeros) == MIN)]
# shorter version
MIN <- min(abs(prob.zeros))
MIN.id <- sek[which(abs(prob.zeros) == MIN)]
# plot
plot(prob.zeros, sek, type="l", pre.plot=grid(),
        col="darkred", bty="n",
        ylab="possible values",
        xlab="finding root ie. zero", main="Optimization problem")
abline(v=MIN, h=MIN.id, col="violetred3", lty=2)
text(MIN, MIN.id+1, pos=3,
        labels=eval(substitute(expression(paste("FUN at ",lambda,
        "=",MIN.id," with value=",MIN)),
        list(MIN.id=MIN.id, MIN=round(MIN,5)))), cex=1, col="blue")
```

Aquí los resultados (véase tambien Fig. 6.116):

```
> # results
> sm.zero
[1] -0.0001124716
> gr.zero
[1] 0.002184457
> MIN
[1] 0.0001124716
> MIN.id
[1] -0.371
```

El propio Jaynes (1963, p.193) describió las ecuaciones como las mejores dada la información disponible: „We are not entitled to assert that the predictions must be "right", only that to make any better ones, we should need more information than was given". Ahora examinamos las distribuciones de probabilidad de las caras del cubo para diferentes valores esperados, para compararlas entre sí (véase la Fig. 6.117). Para ello se utiliza `Jaynes.dice()`. Esta función R espera el valor esperado objetivo *B*, el número de caras del dado *k*

y un valor inicial para el algoritmo de optimización. Utilizando `lapply()`, se obtienen varios valores esperados uno tras otro. Primero el ejemplo original de Jaynes con el valor esperado 4.5 del cubo.

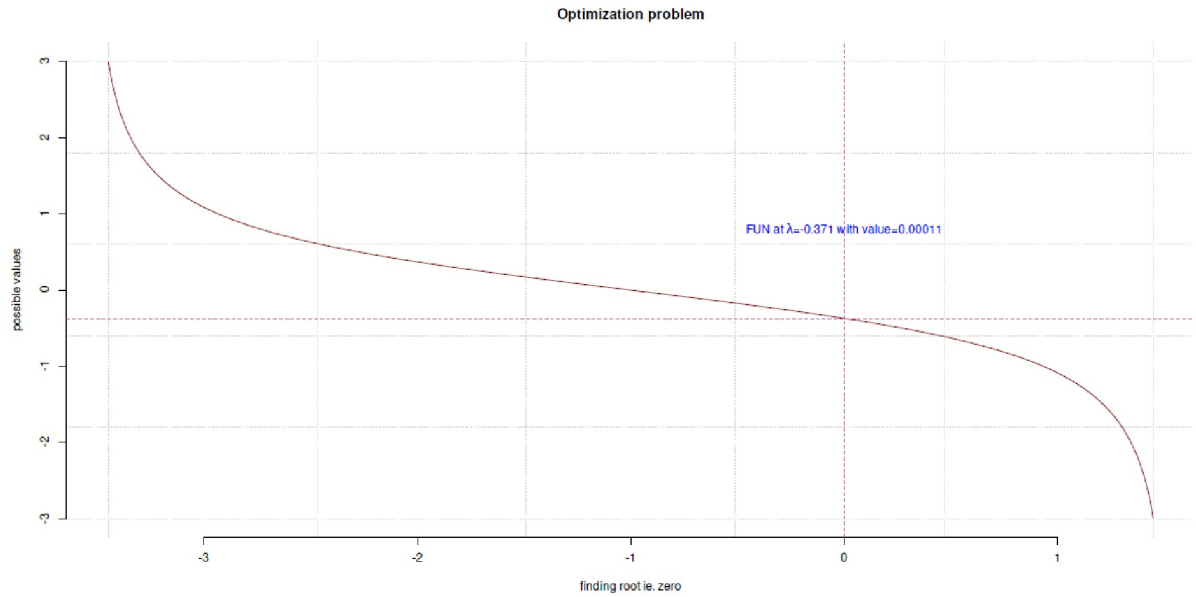


Figura 6.116. *Entropía máxima según Jaynes (1963, optimización manual)*

```
> # example from Jaynes
> Jaynes.dice(B=4.5, k=1:6)$probs
[1] 0.054353 0.078771 0.114160 0.165447 0.239775 0.347494
# fair dice
> Jaynes.dice(B=3.5, k=1:6)$probs
[1] 0.16667 0.16667 0.16667 0.16667 0.16667 0.16667
```

Ahora la aplicación para varios valores esperados:

```
# perform Jaynes' dice problem
Bs <- c(0.5, 2, 3.5, 4, 4.5, 6)
J.dice.res <- lapply(Bs, function(i) Jaynes.dice(B=i))
str(J.dice.res)
```

Sigue el gráfico en Fig. 6.117:

```
par(oma=c(2,1,4,1), "cex.axis"=1, bty="l", mfrow=c(3,2))
for(i in Bs)
{
plot(k, J.dice.res[[i]]$probs, col="violetred3", type="b", pch=21, pre.plot=grid(),
       bg="darkred", main=paste("B=",i,sep=""), bty="n",
       xlab="n-th face of the dice", ylab="MAXENT(p)")
}
mtext("Jaynes' dice problem", outer=TRUE, line=1, cex=1.5, side=3)
```

Como se puede ver, la distribución de probabilidad de las caras del dado en el caso del dado justo con $B = 3.5$ corresponde a una distribución uniforme. Esto no es sorprendente. Si se aplica $B < 3.5$, entonces las caras con menos huecos muestran mayores probabilidades, y para $B > 3.5$ ocurre lo contrario. Las caras del cubo con más rebajes muestran entonces mayores probabilidades. Para el caso de interés B = 4.5, hay una clara desviación de la distribución uniforme, de modo que este cubo difícilmente "pasa" por justo.

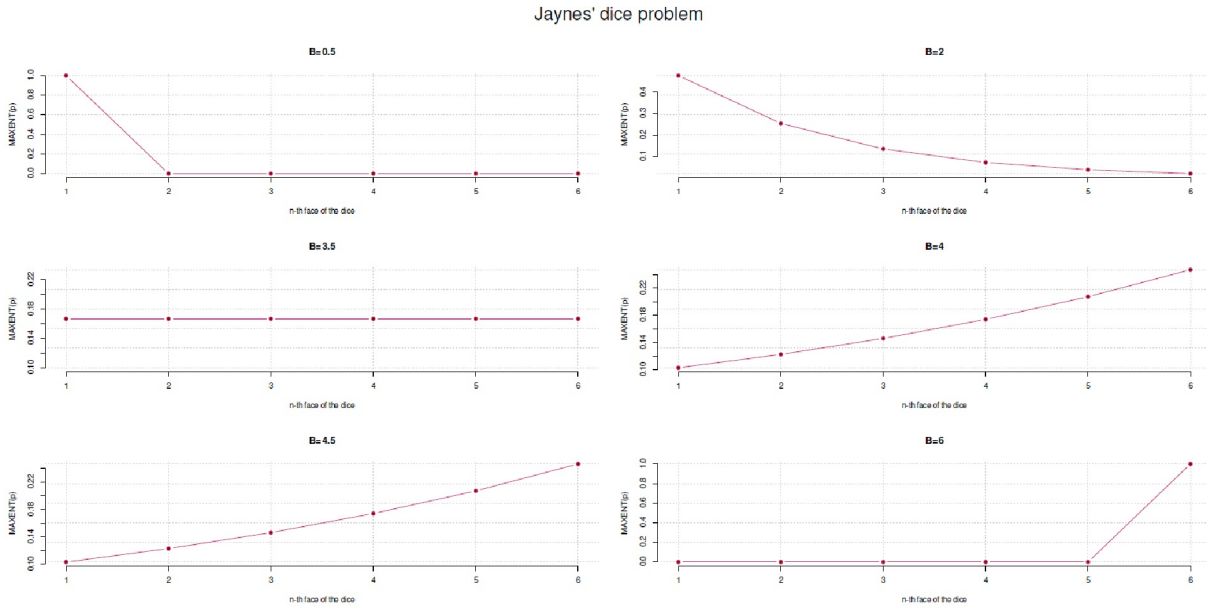


Figura 6.117. *Máxima entropía según Jaynes (1963, optimización sobre diferentes valores esperados)*

En este punto, sería fácil hacer afirmaciones adicionales sobre la forma en que el centro de gravedad del cubo se ha desplazado hacia qué lados y cuáles son las consecuencias en vista de las probabilidades cambiadas. Una discusión del problema del cubo desde el punto de vista actual dan van Enk (2014-08-28) y Mohammad-Djafari (2012-08-28). Una visión crítica de las afirmaciones de Máxima Entropía se encuentra en Rowlinson (1970).

6.14.6 Utilización de información cualitativa para una Prior

Explicar el conocimiento tácito es ante todo un proceso cualitativo. Se trata de reconstruir los conocimientos existentes y ponerlos en una forma lo más libre posible de contradicciones. La traducción de este conocimiento en una distribución a priori es descrita exhaustivamente por O'Hagan, Buck, Daneshkhah, Eiser, Garthwaite, Jenkinson, Oakley y Rakow (2006) en su trabajo. Lamentablemente, no abordan directamente el proceso real de reconstrucción cualitativa, que se debería controlar metodológicamente. En su lugar, estes autores se refieren inmediatamente a las posibilidades de modelización (matemática) del conocimiento previo para derivar una distribución matemática a partir del conocimiento recopilado. Además de numerosos ejemplos y estudios de disciplinas muy diferentes, los autores abordan también los diversos obstáculos (por ejemplo, los efectos sociopsicológicos, tendencias perceptivas inconscientes que pueden distorsionar y estrechar la explicación del conocimiento, etc.) que pueden surgir típicamente al traducir el conocimiento cualitativo de los expertos en hipótesis numéricas sobre una Prior. Se discuten estrategias para tratar distintas cuestiones (por ejemplo, en psicología, economía, etc.). Los autores enumeran innumerables listas de comprobación y procedimientos para trabajar con expertos en diferentes contextos y ordenados por objetivos (por ejemplo, Priores para proporciones, Priores para coeficientes de regresión, etc.) para hacer una suposición sensible al contexto y razonada sobre una prioridad. Después de leer este libro extremadamente cuidadoso e importante, las discusiones sobre la subjetividad de las probabilidades deberían realmente archivarse. Obviamente, no se trata de eso. También debería quedar claro que en casi todas las disciplinas hay tanto conocimiento experto implícito que debería estar justificado recurrir a una Prior uniforme cuando es casi seguro que se dispone de información cuidadosamente justificada para una distribución a priori.

Por un lado, Bayes no puede prescindir de la Prior. Por otro lado, a menudo se encuentran opiniones en libros y artículos que, en última instancia, se refieren siempre a Priores no informativas y recomiendan, por ejemplo, distribuciones uniformes o fijar la masa de la Prior a 0. Consideramos que este enfoque es problemático a menos que exista una justificación sustantiva clara. Como ya se ha explicado, *no existen* Priores no informativos, sino sólo hipótesis que se centran en valores próximos a 0 o en todos los valores simultáneamente. "Informados" en el sentido de tomar una perspectiva concreta son todos los Priores, sin excepción. Una decisión *a favor* de algo es siempre una decisión *en contra* de todo lo demás (Studer, 1998). Tanto si se utiliza la información o no, la información siempre está echada. Siguiendo a Karl Marx (1818-1883) y Jean-Paul Sartre (1905-1980) con sus observaciones (Marx, 1851/1964; Sartre, 1964) sobre la relación dialéctica entre las circunstancias y la acción individual, esto queda más claro y por eso profundizamos en ello. Según Studer (1998 p.32; véase también Gürtler, Studer & Scholz, 2012, nota 66), la cuestión de la relación entre contexto y acción o persona conduce a la pregunta,

"¿Qué hacen las personas de lo que las circunstancias y las condiciones han hecho de ellas?"

Sirven de guía las ideas de Ulrich Oevermann y sus colegas sobre la metodología de la hermenéutica objetiva (Oevermann, Allert, Konau & Krambeck, 1979a; Oevermann, 1981), sobre la profesionalización (Oevermann, 1996b, 1998b), sobre la reconstrucción de casos (Oevermann, 2000) y sobre la dialéctica de lo general y lo particular. En este sentido, es fundamental la idea de que la práctica vital siempre se conceptualiza como una *unidad contradictoria entre la coerción de decidir y la obligación de justificar*, es decir, entre la crisis y la rutina, que es uno de los puntos clave del trabajo de Oevermann y sus colegas. Aplicado a la reconstrucción del conocimiento tácito (experto), esto significa que *antes* de que se puede formular matemáticamente una distribución matemática de cualquier tipo, se debe reconstruir el conocimiento *de forma controlada metodológicamente*. Omitir información es tan inconveniente como distorsionar o falsificar información, ya que entonces cada Prior se basa en información incierta, pero cada una de estas opciones implica una decisión a favor de algo y en contra de todo lo demás. Desde un punto de vista técnico, esto llevaría a un aumento innecesario del componente de ruido en la información o introducir información inadmisible. Ahora bien, se podría decir que, en muchos casos, la masa de los datos hace que la influencia de la información a priori sea insignificante. Sin embargo, esa no es la cuestión. Se trata de una reconstrucción con base científica para llegar a una Prior que explote toda la información disponible. Como método de elección proponemos el análisis de secuencias a partir de la metodología de la hermenéutica objetiva (véase el capítulo 11.5), que permite reconstruir los casos, es decir, responder a la pregunta de qué es lo que realmente está en juego. En relación con una Prior, esto significa dos cosas:

1. ¿Qué información desempeña realmente un papel para una Prior y pertenece a ella?
2. ¿Cómo se debe relacionar esta información individual para crear una estructura bien fundada y, por tanto, formar así una base sustancial para una Prior?

Desde nuestro punto de vista, el análisis de secuencias puede responder a ambas preguntas por completo y en todos los contextos. El esfuerzo no es pequeño, pero al final se obtiene una reconstrucción estructural del caso, de qué información contextual se dispone y cómo está conectada estructuralmente. El siguiente paso es, entonces, junto con los profesionales de las matemáticas, utilizar esto – de nuevo metódicamente controlada, es decir, estructuralmente justificada – como base para una distribución de probabilidad a priori, que entra en el teorema de Bayes como una Prior informada. Por supuesto, es pragmático escalar el procedimiento para grandes cantidades de datos o muchas estimaciones de parámetros, cada una de las cuales requiere una Prior, de tal forma que uno no tenga que pasarse para siempre reconstruyendo la Prior. Por ejemplo, se podría combinar la información y analizarla conjunta-mente. Sin embargo, esto puede no funcionar. A menudo puede ocurrir, y ocurrirá, que con modelos muy complejos parezca casi imposible recopilar tanta información como para que se produzca una reconstrucción sólida para cada parámetro. Se puede recurrir a otras posibilidades como Máxima Entropía (s. cap. 6.14) y elegir Priores con la menor información posible. Entonces se pierden las ventajas de la máxima entropía. La dificultad estriba en que a

menudo apenas será posible contrastar la elección de la Prior según uno u otro procedimiento para estar seguro de haber hecho la elección correcta. Lo que en teoría parece sencillo ("recopilar información contextual, modelizarla matemáticamente y luego maximizar la entropía") a veces la práctica puede resultar a veces extremadamente exigente o incluso demasiado difícil. Entonces, la elección de una Prior muy débil es un paso legítimo.

Sin embargo, cuanto más pequeña es la muestra, más importante resulta este paso de reconstrucción cualitativa. Esto es probablemente cierto para muchos experimentos psicológicos que tradicionalmente trabajan con muestras pequeñas; y aquí no se debe escatimar el esfuerzo.

Siempre que el conocimiento experto pueda explicarse científicamente, se debería utilizarlo, ya que es precisamente el *fallo potencial inherente* a una Prior informativa lo que hace avanzar el aprendizaje y el conocimiento. Esto sólo ocurre parcialmente, si es que ocurre, cuando se utiliza una distribución uniforme o una Prior de Jeffreys, que no tiene nada que ver con el contexto. La distribución uniforme no prefiere nada y la Prior de Jeffreys con Beta(0.5, 0.5) prefiere los extremos, pero no tanto como la Prior de Haldane (Haldane, 1932) y da poca o ninguna importancia al intervalo intermedio. Es cuestionable que tales especificaciones puedan resistir en la práctica un examen sustantivo del conocimiento contextual. Vemos una excepción en el enfoque de entropía máxima, ya que este principio junto con la segunda ley de la termodinámica tiene un alto contenido de realidad en el plano de la información. Además, el enfoque permite incluir información a priori.

O'Hagan, Buck, Daneshkhah, Eiser, Garthwaite, Jenkinson, Oakley y Rakow (2006) piden explícitamente la programación de un software especial a la hora de crear Priores. Esto debería ayudar a los expertos a poner en práctica sus ideas en tiempo real y a dar salida a los supuestos realizados tanto gráficamente como numéricamente, permitiendo así una verificación visual inmediata. Los métodos habituales consisten, sencillamente, en pedir a los expertos que proporcionen valores para la tendencia central (mediana) y los cuantiles u otros rangos de valores adecuados. Otra instrucción consiste en realizar estos supuestos de forma gráfica (por ejemplo, en papel) o física, pidiendo a los expertos que distribuyan un determinado conjunto limitado de elementos (por ejemplo, bloques) a lo largo de la gama de valores posibles. Estos elementos representan el estado actual de los conocimientos sobre un problema. Visualizan a grandes rasgos la forma de la distribución. Los estadísticos derivan entonces una distribución con los parámetros correspondientes. En cualquier caso, las ideas parametrizadas deben devolverse a los expertos para que las confirmen y, si es necesario, para discutir si es necesario introducir cambios. Debido a los problemas de optimización matemática, es posible que las hipótesis de distribución no siempre se correspondan exactamente con las de los expertos. Esto no es necesariamente una desventaja, ya que permite discutir la solución encontrada en términos de su viabilidad. Además, en realidad es necesario crear un vínculo entre los estadísticos y los expertos que, como ya se ha explicado, en primer lugar ayude a los expertos a obtener sus conocimientos de forma controlada y a darle una forma correcta en términos de contenido, o a condensar el contenido expresado precisamente en una hipótesis de estructura de caso. En este sentido, los instrumentos del *paradigma de codificación* y el *análisis secuencial* presentados en los capítulos 9 y 11 prometen avances sustanciales, ya que entonces el conocimiento no se expresa arbitrariamente, sino que se explica y resume a lo largo de un proceso científico, que repercute en la posterior transformación en forma de elementos estadísticos. La bibliografía actual sobre la explicación del conocimiento tácito no menciona este enfoque cualitativo metodológicamente controlado, lo que sugiere que la etapa de reconstrucción cualitativa se abandona más o menos en favor del trabajo con distribuciones matemáticas o no se lleva a cabo de forma metodológicamente controlada.

Tras las explicaciones de la teoría de Bayes, se presentan tanto el desarrollo de una Prior como el proceso de actualización en el curso del teorema de Bayes sobre la base de un estudio de caso muy detallado sobre las *tasas de aprobados en la terapia hospitalaria de la adicción* (véase el capítulo 6.3.2 para un estudio de caso discreto). La exhaustividad de la discusión cualitativa tiene como objetivo explicar las consideraciones que no suelen encontrarse en artículos y libros de texto sobre cómo se puede llegar realmente a supuestos sobre una Prior. A esto le siguen varios estudios de caso empíricos más pequeños sobre el análisis bayesiano:

- Comenzamos con la comparación de las longitudes corporales de los presidentes estadounidenses y sus contrincantes (véase el capítulo 6.15.1).

- A continuación, se analiza la evolución de los índices de aprobados en la terapia de adicción con hospitalización a lo largo de varios años (véase el capítulo 6.15.2, véase también el capítulo 6.8.1.2).
- Continuando con el estudio de caso sobre la autopromoción presidencial del capítulo 6.14.3, se aplican diferentes métodos de análisis bayesiano a los mismos datos y se comparan los resultados (véase el capítulo 6.15.3).

6.15 Casos prácticos bayesianos

6.15.1 Alturas de presidentes

Un ejemplo popular de datos para explorar las diferencias entre dos grupos son las alturas de los presidentes estadounidenses y sus contrincantes, naturalmente en la línea de si *los más altos* son también los que ganaron. Tomamos los datos de de la entrada correspondiente de Wikipedia (2019g) y esperamos que no hayan sido *embellecidos*. Una comparación aleatoria con otros sitios web (por ejemplo, Gal & Lee, 2019; AmericanPresidents, 2019) ya muestra una diferencia de 3 cm para Donald Trump (Gabbat, 2018; Informe oficial de la Casa Blanca de Jackson, 2018). George W. Bush también aparece en Wikipedia (2019g) con una altura de 1,7 cm superior a la de AmericanPresidents (2019), pero no difiere a Gal y Lee(2019). Hay acuerdo, sin embargo, en que Abraham Lincoln fue el presidente estadounidense más alto, con 1,93 m. A continuación, utilizamos los datos de Wikipedia para demostrar el enfoque, pero lo consideramos con cautela. Aparte de eso, sería legítimo preguntarse cuál es el propósito más profundo de investigar la estatura de los presidentes estadounidenses y sus contrincantes. Preferimos no responder a esa pregunta. Pero tales comparaciones se pueden encontrar y ellas constituyen un buen ejemplo de análisis de datos.

Dado que diferentes personas (presidentes, candidatos) aparecen varias veces debido a la repetición de las elecciones, es necesario aclarar la pregunta, lo que a su vez afecta a la definición exacta de la unidad de estudio. Se puede tratar, de casos o de personas y, en consecuencia, se pregunta por encuentros y resultados electorales o por comparaciones de grupos de presidentes y candidatos.

Casos – En cada elección hay un presidente en ejercicio y un aspirante o dos aspirantes, pero siempre un único ganador. Una diferencia en las alturas de estas dos personas nos dice entonces algo sobre la comparación directa de estas dos personas en un momento dado. Aunque esto corresponde a una comparación de grupo a lo largo de todas las elecciones, se trata, por otro lado, de datos dependientes (= comparaciones por pares) o de pruebas múltiples (por ejemplo, primero como aspirante, luego como presidente en ejercicio). Por lo tanto, sólo se pueden analizar las diferencias de estas comparaciones por pares con respecto a una hipótesis de interés. Las comparaciones múltiples no plantean problemas, ya que el análisis se basa en los emparejamientos respectivos por año electoral. Todos los datos se incluyen en el cálculo.

Personas – **A**quí es importante eliminar todas las entradas múltiples y trabajar sólo con los tamaños corporales. El análisis de las diferencias en el tamaño corporal pregunta por las diferencias medias en el grupo de presidentes (= ganadores de las elecciones) o aspirantes. La comparación se realiza indepen-dientemente de las respectivas elecciones y encuentros personales. Se trata de una comparación de grupos independientes.

En sentido estricto, habría que introducir un factor de corrección, ya que las mujeres tienden a ser ligeramente más pequeñas que los hombres. Sin embargo, dado que, con la excepción de Hillary Clinton, ninguna mujer ha sido nunca candidata o presidenta, ignoramos esto en aras de la simplicidad y aceptamos la distorsión asociada. Sería demasiado complejo introducir ese factor de corrección – ¿y según qué criterios?

Desde el punto de vista estadístico, este problema puede resolverse utilizando la prueba *t* para grupos dependientes (= casos) o independientes (= personas). El enfoque bayesiano tiene diversas variantes, como `bayes.t.test()` del paquete `Bolstad` de R, `ttestBF()` de `BayesFactor` (con factor Bayes) o

`bayes.t.test()` de `BayesianFirstAid`. Además, existe `BESTmcmc()` del paquete R `BEST`, que se basa en el artículo de Kruschke (2013a) y que implementa la prueba de dos grupos con JAGS mediante MCMC.

Lejos de los paquetes R, existen adaptaciones de Mathematica™ (Studer, 1998 y Gregory, 2006), que un autor (Gürtler) portó a R y que se basan en el trabajo seminal de Bretthorst (1993). Bretthorst elaboró una solución exacta y analítica al problema Behrens-Fisher (Behrens, 1929; Fisher, 1935), es decir, la cuestión de las medias y varianzas desiguales. El script R de la adaptación de Gregory (2006) produce opcionalmente gráficos para las probabilidades posteriores. En la versión de Studer (1998), esto se ha ampliado para incluir una función para comparar dos tasas de éxito.

En primer lugar, se leen los datos ya extendidos a los datos originales. Así, existen adicionalmente las columnas `h.diff.TF` (=valores verdaderos, si el ganador fue mayor que el retador), `h.diff` (= diferencia en cm del ganador frente al retador) así como `WP` y `OP`, que corresponden respectivamente a la combinación de ganador o contrincante y la afiliación partidista respectiva (es decir, afiliación Republicana o Demócrata) (`ptII_quan_Bayes_case_presidential-heights.r`).

```
# read enhanced file
pres.h <- read.table("presidential-heights.tab", sep="\t", header=TRUE)
dim(pres.h)
str(pres.h)
head(pres.h)
# add diffs
pres.h <- data.frame(pres.h, h.diff=with(pres.h, W.cm-O.cm))
```

Hay datos que faltan, ¿dónde están? No los tratamos aquí (véase el capítulo 4.4.8).

```
# remove NAs
data.frame("na"=apply(pres.h,2,function(i) sum(is.na(i))))
naids <- which(is.na(pres.h),arr.ind=TRUE)
naids
naids.unique <- unique(naids[,"row"])
naids.unique
pres.h.nona <- pres.h[-naids.unique,]
pres.h.nona$h.diff.TF <- with(pres.h.nona, h.diff > 0)
head(pres.h.nona)
```

Ahora hay que identificar las entradas múltiples de los ganadores y los aspirantes. Los datos se extraen y se escriben en una lista.

```
# identify appearances of persons more than once R-Code
# unless we want to make the diff for each election
# > > > then these are cases of winner vs. main.opponent < < <
W.double <- sort(table(pres.h.nona[,"winner"]) >1, dec=TRUE)
O.double <- sort(table(pres.h.nona[,"main.opponent"]) >1, dec=TRUE)
W.double[W.double]
O.double[O.double]
# use persons, not cases and reduce NAs 'en passent'
W.cm.RED <- data.frame(do.call("rbind",
        strsplit(unique(with(pres.h.nona,
        paste(winner,W.cm,sep=":"))),":")))
O.cm.RED <- data.frame(do.call("rbind",
        strsplit(unique(with(pres.h.nona,
        paste(main.opponent,O.cm,sep=":"))),":")))
colnames(W.cm.RED) <- c("winner","W.cm")
colnames(O.cm.RED) <- c("opponent","O.cm")
W.cm.RED[,"W.cm"] <- as.numeric(as.character(W.cm.RED[,"W.cm"]))
O.cm.RED[,"O.cm"] <- as.numeric(as.character(O.cm.RED[,"O.cm"]))
W.cm.RED <- W.cm.RED[,"W.cm"][!is.na(W.cm.RED[,"W.cm"])]
O.cm.RED <- O.cm.RED[,"O.cm"][!is.na(O.cm.RED[,"O.cm"])]
W.cm.RED
```

```
O.cm.RED
# combine - cases and persons
dats <- list(W=pres.h.nona$W.cm, O=pres.h.nona$O.cm,
        W.red=W.cm.RED,
        O.red=O.cm.RED)
dats
```

A continuación se presentan las estadísticas descriptivas. Para una mejor comprensión de los datos, a continuación se presenta un histograma o estimaciones de densidad para las diferencias (véase la Fig. 6.118), así como gráficos de caja para los dos conjuntos de datos (casos, personas, véase la Fig. 6.119).

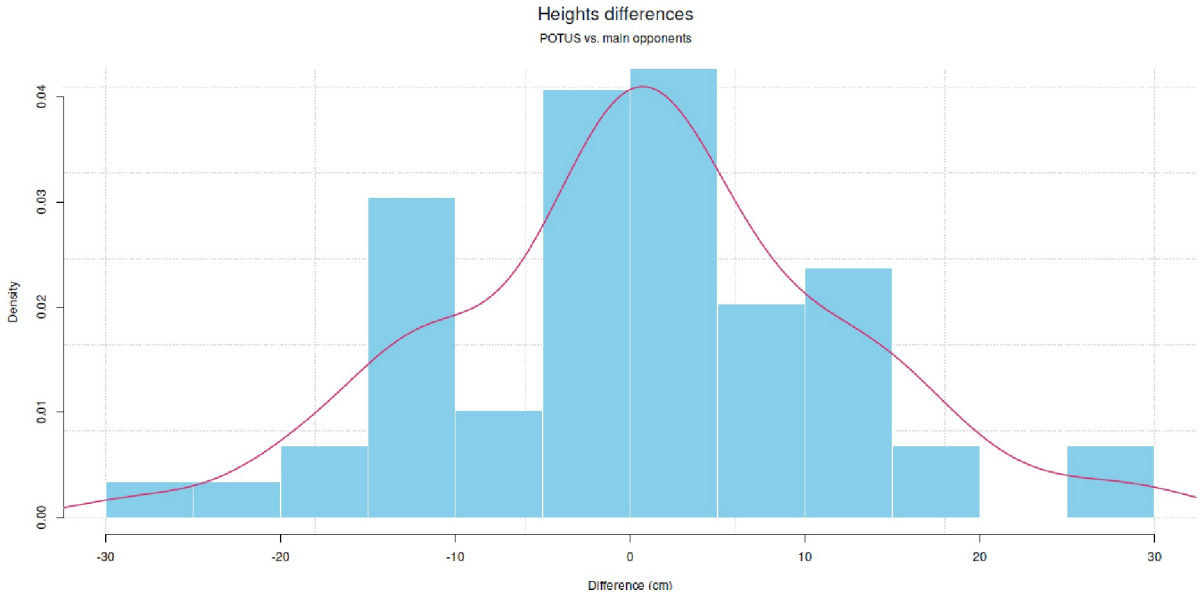


Figura 6.118. *Presidentes y aspirantes de EE.UU.*
(tamaños corporales, histograma y estimación de densidad)

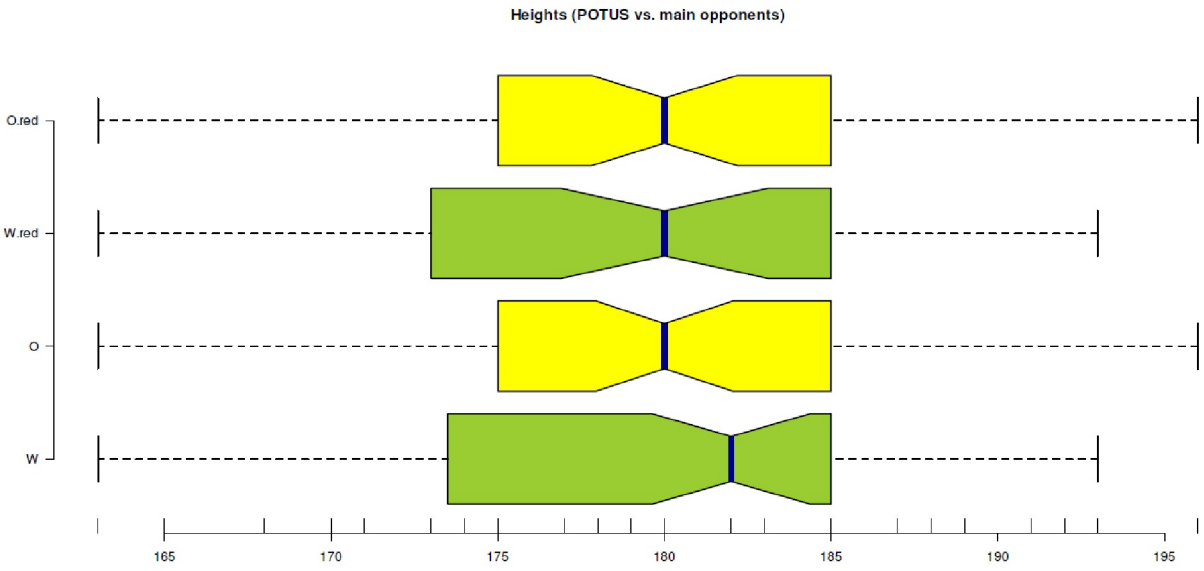


Figura 6.119. *Presidentes de EE.UU. y aspirantes (tamaños corporales, boxplots)*

```
# plot differences R-Code
h.dens <- density(pres.h.nona$h.diff, na.rm=TRUE)
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l")
hist(pres.h.nona$h.diff, prob=TRUE, xlab="Difference (cm)",
       ylab="Density", main="",
       breaks=20, col="skyblue", border="white",
       ylim=c(0,max(h.dens$y)), pre.plot=grid())
lines(h.dens, col="violetred3", lwd=2, lty=1)
mtext("Heights differences", outer=TRUE, line=-1, cex=1.5, side=3)
mtext("POTUS vs. main opponents", outer=TRUE, line=-2.5, cex=1, side=3)
# boxplots
boxplot(dats, col=c("yellowgreen","yellow"), notch=TRUE,
       main="Heights (POTUS vs. main opponents)",
       las=1, frame.plot=FALSE, horizontal=TRUE,
       outly=2, lwd=2, medcol="darkblue")
rug(unlist(dats))
```

Las variables `W` y `O` incluyen los registros no reducidos con respuestas múltiples, mientras que `W.red` y `O.red` no incluyen respuestas múltiples.

```
> # descriptive statistics
> summary(pres.h.nona$h.diff)
 Min.     1st Qu. Median Mean   3rd Qu. Max.
-28.0000 -5.0000  1.0000 0.7797 8.0000  30.0000
> dats.desc <- do.call("rbind", lapply(dats,
       + function(x) c(c(summary(x),VAR=var(x),SD=sd(x),))) )
> dats.desc
      Min. 1st Qu. Median Mean     3rd Qu. Max. VAR      SD
W     163  173.5   182    180.3051 185     193  59.38808 7.706366
O     163  175.0   180    179.5254 185     196  49.18469 7.013180
W.red 163  173.0   180    179.6486 185     193  55.06757 7.420752
O.red 163  175.0   180    179.5385 185     196  53.46908 7.312255
```

Alternativamente, antes de examinar los datos con más detalle mediante una prueba *t*, una pequeña tabla puede dar una primera indicación.

```
> # who is taller?
> h.diff.TF.tab <- table(pres.h.nona$h.diff.TF)
> h.diff.TF.tab
FALSE TRUE
28    31
> h.W.successes <- h.diff.TF.tab["TRUE"]
> h.N <- sum(h.diff.TF.tab)
```

La cuestión de si el ganador o el contrincante fue mayor puede examinarse de forma frecuentista o bayesiana mediante una prueba binomial. Para la variante bayesiana utilizamos `bayes.binom.test()` del paquete `BayesianFirstAid` de R. Comenzamos con la prueba clásica

```
> # classic
> binom.test(x=h.W.successes, n=h.N, p=0.5, alternative="greater")
Exact binomial test
data: h.W.successes and h.N
number of successes = 31, number of trials = 59, p-value = 0.3974
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
0.4109916 1.0000000
sample estimates:
probability of success
0.5254237
```

y comparamos el resultado con la prueba binomial bayesiana:

```
> # Bayes
> h.bbt <- bayes.binom.test(x=h.W.successes, n=h.N,
        + comp.theta=0.5, cred.mass=0.87)
> summary(h.bbt)
Data
number of successes = 31, number of trials = 59
Model parameters and generated quantities
theta: the relative frequency of success
x_pred: predicted number of successes in a replication
Measures
       mean   sd    HDIlo HDIup  %<comp %>comp
theta  0.525  0.064 0.43   0.624 0.349  0.651
x_pred 30.981 5.366 22.00 38.000 0.000  1.000
'HDIlo' and 'HDIup' are the limits of a 87% HDI credible interval.
'%<comp' and '%>comp' are the probabilities of the respective
parameter being smaller or larger than 0.5.
Quantiles
       q2.5%  q25%   median q75%   q97.5%
theta   0.398  0.481  0.525  0.568  0.647
x_pred 20.000 27.000 31.000 35.000 41.000
```

Básicamente, ambas pruebas conducen a la misma decisión, aunque la prueba bayesiana mide naturalmente la incertidumbre y la prueba clásica pertenece al NHST y dice poco en caso de resultado no significativo. A continuación, realizamos un diagnóstico bayesiano de la Posterior y simulaciones MCMC (salida no impresa)

```
plot(h.bbt)
lapply(h.bbt$mcmc_samples, hdi, credMass=0.87)
diagnostics(h.bbt)
```

o miramos la llamada al código JAGS

```
> model.code(h.bbt)
require(rjags)
# Setting up the data
x <- 31
n <- 59
# The model string written in the JAGS language
model_string <- "model {
x ~ dbinom(theta, n)
theta ~ dbeta(1, 1)
x_pred ~ dbinom(theta, n)
}"
# Running the model
model <- jags.model(textConnection(model_string),
        data = list(x = x, n = n),
        n.chains = 3, n.adapt=1000)
samples <- coda.samples(model, c("theta", "x_pred"), n.iter=5000)
# Inspecting the posterior
plot(samples)
summary(samples)
```

Mientras que la prueba frecuentista sólo puede rechazar o mantener la hipótesis nula, la prueba bayesiana cuantifica la incertidumbre en el contexto de la Prior. Como Prior, adoptamos el estándar de `bayes.binom.test()`, una distribución uniforme, es decir, una distribución beta con los parámetros a = 1 y b = 1. Aunque la función no permite modificar la Prior, esto se puede conseguir con relativa facilidad

modificando directamente el código de R y JAGS. Recibimos el código por examinar sucesivamente las funciones en cuestión y sus códigos fuente. A partir de éste, se pueden extraer los lugares para construir en una propia Prior.

```
# want to change prior? R-Code
bayes.binom.test
jags_binom_test
binom_model_string
run_jags
```

Así, resulta que basta con sustituir la función de R `BayesianFirstAid:::jags_binom_test()` por otra propia función. La llamamos `bayes.binom.alt()`. Se puede utilizar otros parámetros para la distribución a priori beta o directamente el código JAGS para cambiar el modelo por completo. `Comp.theta` = 0.5 se elige como valor crítico de comparación. Es decir, se examina la hipótesis de que las alturas corporales se distribuyen por igual. Elegimos los valores beta a = 3 y b = 20 para la Prior.

```
prior.a <- 3 R-Code
prior.b <- 20
bbinom.model.string <- paste(
"model {\n x ~ dbinom(theta, n)\n theta ~ dbeta(", prior.a,", ",prior.b,
       ")\n x_pred ~ dbinom(theta, n)\n}",sep="")
bbinom.model.string
x <- h.W.successes
n <- h.N
x_name <-"W"
n_name <- "elections"
```

La llamada al modelo da nuevos resultados:

```
> # use alternative version of
> # bayes.binom() from BayesianFirstAid
> h.bbt.res <- bayes.binom.alt(x=x, n=n, cred.mass=0.87, comp.theta=0.5,
       + x_name=x_name, n_name=n_name, model_string=bbinom.model.string)
Data
number of successes = 31, number of trials = 59
Model parameters and generated quantities
theta: the relative frequency of success
x_pred: predicted number of successes in a replication
Measures
       mean   sd    HDIlo  HDIup  %<comp %>comp
theta   0.415 0.054  0.332  0.497 0.939  0.061
x_pred 24.469 4.954 16.000 31.000 0.000  1.000
'HDIlo' and 'HDIup' are the limits of a 87% HDI credible interval.
'%<comp' and '%>comp' are the probabilities of the respective
parameter being smaller or larger than 0.5.
Quantiles
       q2.5%  q25%   median q75%   q97.5%
theta   0.311  0.378  0.414  0.451  0.523
x_pred 15.000 21.000 24.000 28.000 34.000
       mean       sd         HDI% comp HDIlo      HDIup
theta   0.4147217 0.05422217 87   0.5   0.3321265  0.4969163
x_pred 24.4689556 4.95388326 87   0.5  16.0000000 31.0000000
       %>comp     %<comp       q2.5%     q25%       median
theta  0.06097507 9.390249e-01  0.310511  0.3778157  0.4137438
x_pred 0.99997778 2.222123e-05 15.000000 21.0000000 24.0000000
       q75%       q97.5%     mcmc_se      Rhat      n_eff
theta   0.4510805  0.5228482 0.0002605189 0.9999796 43381
x_pred 28.0000000 34.0000000 0.0235536569 1.0000018 44255
```

Como alternativa

```
h.bbt.res.alt1 <- bayes.binom.alt(x=x, n=n, cred.mass=0.87,
        comp.theta=0.5, x_name=x_name, n_name=n_name)
h.bbt.res.alt2 <- bayes.binom.alt(x=x, n=n, cred.mass=0.87,
        comp.theta=0.5, x_name=x_name, n_name=n_name, prior.a=2, prior.b=45)
```

Las salidas se ejecutan por defecto con `summary()`, `plot()` y para los datos MCMC mediante `diagnostics()`. La nueva función escrita en R muestra los resultados y un gráfico por defecto. Los HDI se pueden crear mediante una llamada a `lapply()`, la salida gráfica de la Posterior se muestra en la Figura 6.120.

```
# HDI and diagnostics of MCMC chain(s)
lapply(h.bbt$mcmc_samples, hdi, credMass=0.69)
diagnostics(h.bbt.res)
# output model etc.
summary(h.bbt.res)
plot(h.bbt.res)
# initial values
str(h.bbt.res)
h.bbt.res$x
h.bbt.res$n
```

y el modelo JAGS utilizado para las pruebas:

```
> # model
> cat(h.bbt.res$model)
model {
x ~ dbinom(theta, n)
theta ~ dbeta(3, 20)
x_pred ~ dbinom(theta, n)
}
```

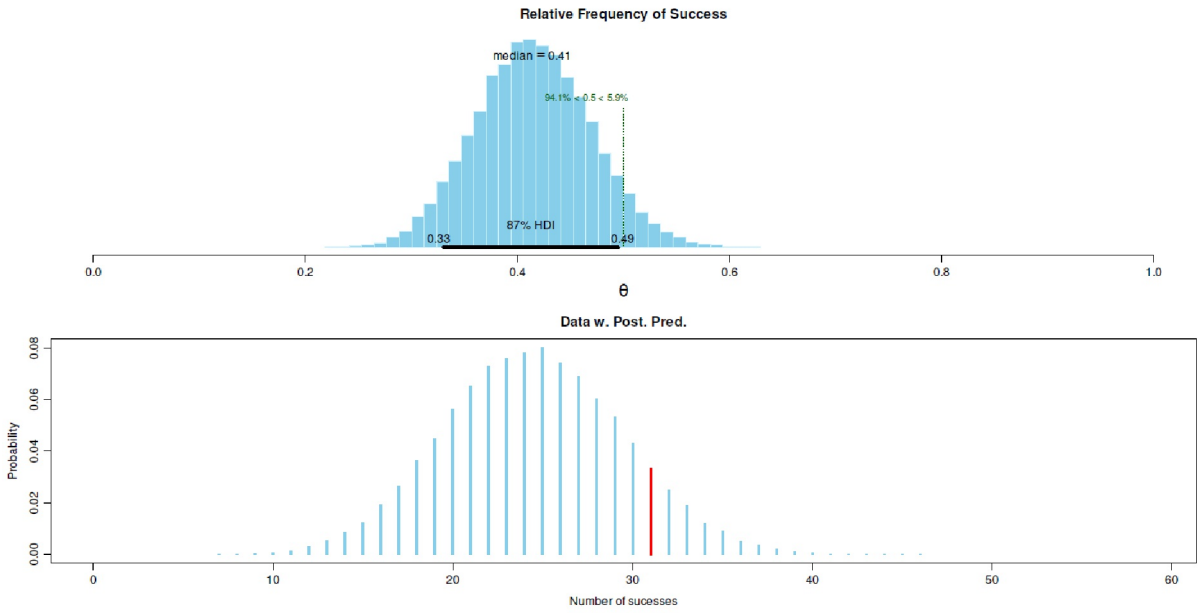


Figura 6.120. *Presidentes de EE.UU. y aspirantes*
(Alturas corporales, binomialtest bayesiano Posterior, Prior informado con a = 3, b = 20)

> **Tarea 6.11: Cambios debidos a la Prior**
>
> La tarea para el lector interesado sería elegir otra Prior, repetir el análisis y comparar los resultados.

Un cambio de la Prior hacía $a = 0.001$ o $b = 0.001$ con un valor crítico de comparación para θ y un HDI del 87 % obtenemos por

```
# change priors
prior.a <- 0.001
prior.b <- 0.001
comp.theta <- 0.5
cred.mass <- 0.87
h.bbt.res.1 <- bayes.binom.alt(x=x, n=n, cred.mass=cred.mass,
        comp.theta=comp.theta,
        prior.a=prior.a, prior.b=prior.b)
lapply(h.bbt.res.1$mcmc_samples, hdi, credMass=0.69)
```

A continuación inspeccionamos el MCMC junto con un análisis gráfico (véase Fig. 6.121):

```
> diagnostics(h.bbt.res)
Iterations = 2:15001
Thinning interval = 1
Number of chains = 3
Sample size per chain = 15000
Diagnostic measures
       mean   sd    mcmc_se n_eff Rhat
theta   0.415 0.054 0.000   43933 1
x_pred 24.444 4.935 0.023   45057 1
mcmc_se: the estimated standard error of the MCMC approximation
of the mean.
n_eff: a crude measure of effective MCMC sample size.
Rhat: the potential scale reduction factor (at convergence, Rhat=1).
Model parameters and generated quantities
theta: The relative frequency of success
x_pred: Predicted number of successes in a replication
```

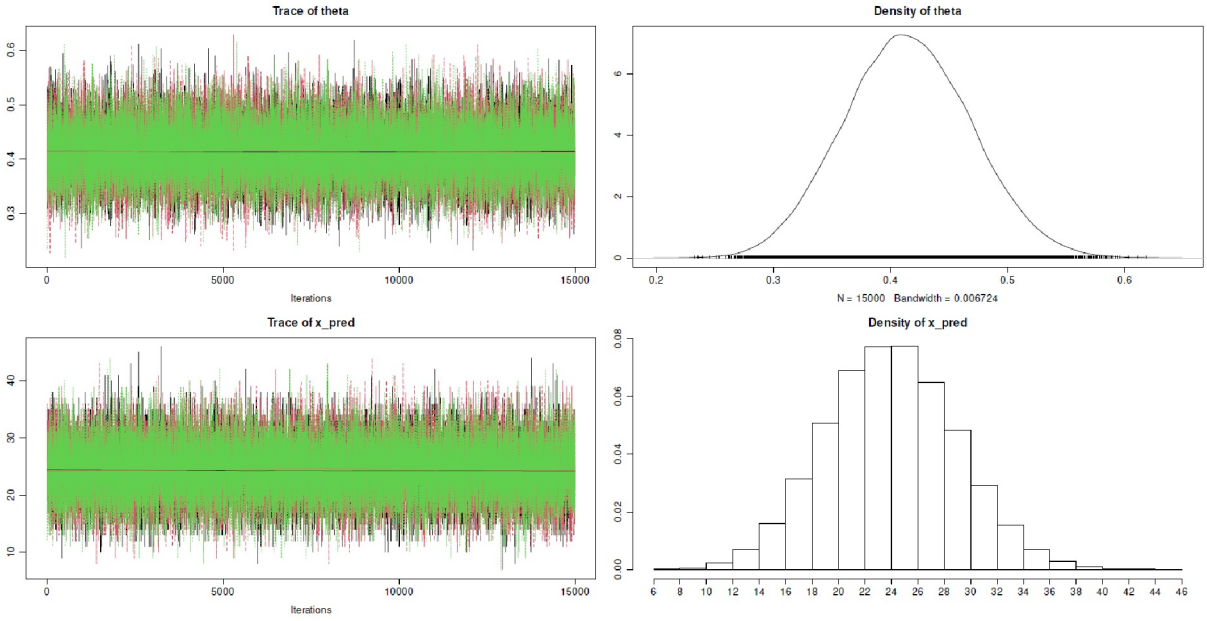


Figura 6.121. Presidentes de EE.UU. y aspirantes
(tamaños corporales, binomialtest bayesiano, diagnóstico MCMC y Posterior, Prior de baja información con a = 0.001, b = 0.001).

De forma muy similar, se podría investigar si existe una dependencia de las diferencias en los tamaños corporales con la afiliación partidista. Esto requiere otra tabla y la prueba `bayes.proportion.test()` del paquete `BayesianFirstAid` de R. Para simplificar, bastaría con tomar sólo las combinaciones de candidatos republicanos y demócratas. Esto requiere una reducción de la tabla de frecuencias. Sería tarea de los lectores realizar estos cálculos.

Tras este trabajo preliminar, el conjunto de datos se puede analizar de forma clásica con la prueba *t* para muestras dependientes y varianzas desiguales más el tamaño del *efecto d de Cohen*. Comenzamos con las elecciones

```
> # test the difference in means
>
> # data base
> # cases / each election
> W <- dats[["W"]]
> O <- dats[["O"]]
>
> # classic
> with(dats, t.test(W, O, alternative="greater", paired=TRUE, var.equal=FALSE))
Paired t-test
data: W and O
t = 0.52748, df = 58, p-value = 0.2999
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
-1.691031 Inf
sample estimates:
mean of the differences
0.779661
> cohensd(W[!is.na(W)], O[!is.na(O)])
 d|mean sd  d|pooled sd
-0.1058182 -0.1058182
```

y comparamos las personas

```
# persons (presidents vs. opponents)
> W.red <- dats[["W.red"]]
> O.red <- dats[["O.red"]]
> # classic
> with(dats, t.test(W.red, O.red, alternative="greater",
       + paired=FALSE, var.equal=FALSE))
Welch Two Sample t-test
data: W.red and O.red
t = 0.069459, df = 76.988, p-value = 0.4724
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
-2.530938 Inf
sample estimates:
mean of x mean of y
179.6486  179.5385
> cohensd(W.red,O.red)
 d|mean sd   d|pooled sd
-0.01495745 -0.01497648
```

El paquete de R `BayesianFirstAid` también tiene una prueba *t* bayesiana. Podemos probar sobre los casos (por elección)

```
# cases / each election
pres.h.bt.res1 <- with(dats, bayes.t.test(W, O, paired=TRUE))
summary(pres.h.bt.res1)
plot(pres.h.bt.res1)
diagnostics(pres.h.bt.res1)
model.code(pres.h.bt.res1)
```

o sobre las personas (más tarde presidente contra aspirante; salidas no impresos):

```
# persons (presidents vs. opponents)
pres.h.bt.res2 <- with(dats, bayes.t.test(W.red, O.red,
paired=FALSE))
# not really different to BEST
summary(pres.h.bt.res2)
plot(pres.h.bt.res2)
diagnostics(pres.h.bt.res2)
model.code(pres.h.bt.res2)
```

A esto le sigue el análisis bayesiano con `BESTmcmc()` – que incluye gráficos, HDI y diagnósticos de las simulaciones MCMC (véase la Fig. 6.122 para una visión general).

```
# very wide prior (default), see ?BESTmcmc R-Code
# cases
W.nona <- pres.h.nona[,"W.cm"]
O.nona <- pres.h.nona[,"O.cm"]
# calculate model
pres.best.res <- BESTmcmc(W.nona, O.nona, rnd.seed=84445)
str(pres.best.res)
names(pres.best.res)
# plot and summaries
plot(pres.best.res)
summary(pres.best.res)
print(pres.best.res)
plotPost(pres.best.res[,"mu1"])
pairs.BEST(pres.best.res)
plotAll(pres.best.res)
```

```
plotPostPred(pres.best.res)
hdi(pres.best.res)
```

Por defecto, `BESTmcmc()` aplica valores de Priores muy amplios, véase `?BESTmcmc`. El resumen de la estimación recibimos por `summary()` y `print()`:

```
> summary(pres.best.res)
          mean   median mode   HDI% HDIlo  HDIup   compVal %>compVal
mu1       180.39 180.39 180.42 95   178.34 182.42
mu2       179.53 179.53 179.54 95   177.72 181.41
muDiff      0.86   0.86   0.85 95  -1.94     3.57  0       73.1
sigma1      7.70   7.65   7.58 95   6.25     9.24
sigma2      6.97   6.92   6.87 95   5.65     8.41
sigmaDiff   0.74   0.73   0.69 95  -1.307    2.77  0       76.5
nu         45.10  36.60  22.33 95   4.92   108.55
log10nu     1.55   1.56   1.62 95   0.97     2.13
effSz       0.12   0.12   0.12 95  -0.26     0.49  0       73.1
> print(pres.best.res)
MCMC fit results for BEST analysis:
100002 simulations saved.
     mean    sd   median HDIlo  HDIup Rhat n.eff
mu1  180.39  1.04 180.39 178.34 182.42 1   60630
mu2  179.53  0.94 179.53 177.72 181.41 1   60541
nu    45.10 32.02  36.60   4.92 108.55 1   22754
sigma1 7.70  0.77   7.65   6.25   9.24 1   50630
sigma2 6.97  0.71   6.92   5.65   8.41 1   49293
'HDIlo' and 'HDIup' are the limits of a 95% HDI credible interval.
'Rhat' is the potential scale reduction factor (at convergence, Rhat=1).
'n.eff' is a crude measure of effective sample size.
```

Se puede utilizar la Posterior para comprobar directamente hipótesis, por ejemplo, la probabilidad de que la diferencia sea mayor o menor que 0 o mayor o menor que 3.

```
> # diff and ROPE
> meanDiff.h <- pres.best.res$mu1 - pres.best.res$mu2
> mean(meanDiff.h > 0)
[1] 0.7308254
> mean(meanDiff.h < 0)
[1] 0.2691746
> mean(meanDiff.h > 3)
[1] 0.06283874
> mean(meanDiff.h < 3)
[1] 0.9371613
```

A partir de ahí se puede obtener una Odds ratio.

```
> # BF_greater-vs-smaller(winner)
> mean(meanDiff.h > 0) / mean(meanDiff.h < 0)
[1] 2.715061
> mean(meanDiff.h < 0) / mean(meanDiff.h > 0)
[1] 0.3683159
```

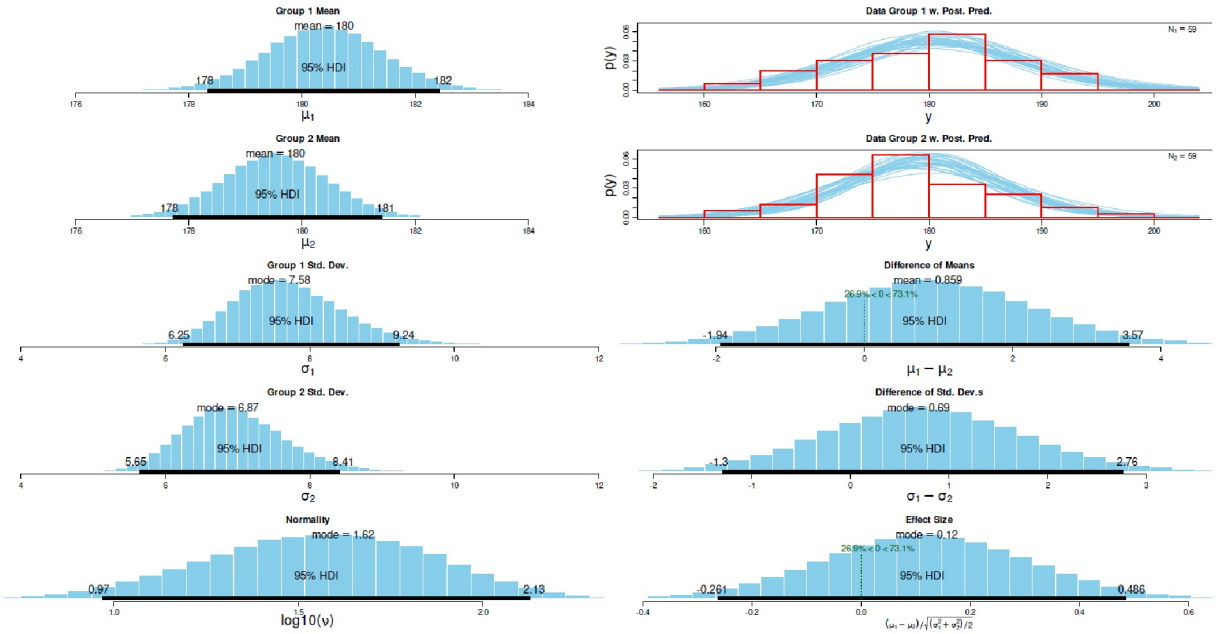


Figura 6.122. *Presidentes de EE.UU. y aspirantes*
(alturas, prueba t bayesiana con BESTmcmc(), Posterior)

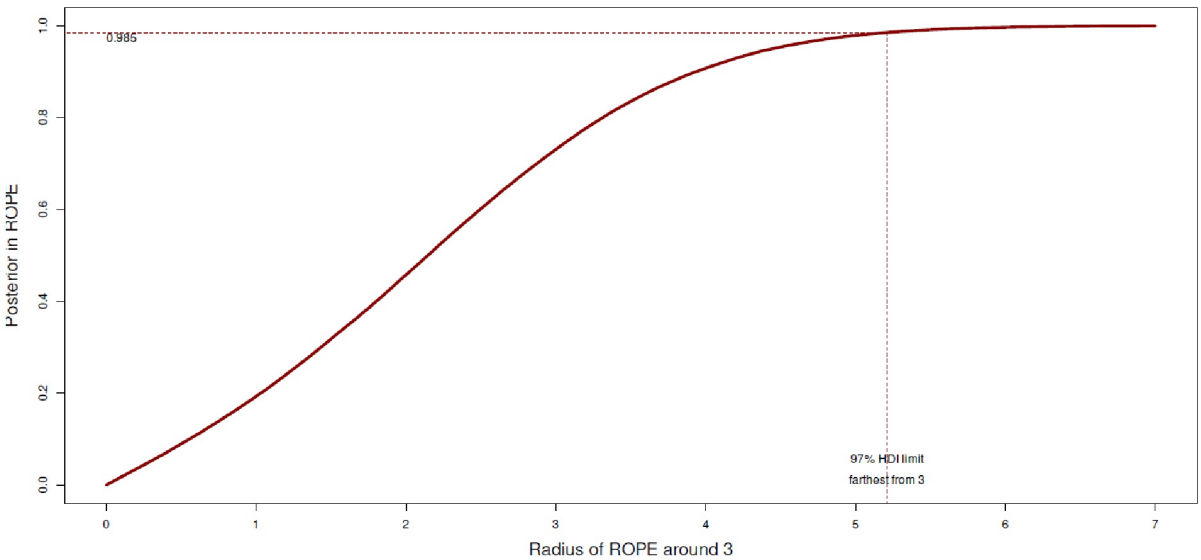


Figura 6.123. *Presidentes de EE.UU. y aspirantes*
(alturas, prueba t bayesiana con BESTmcmc(), ROPE)

Y, por último, trazamos el radio alrededor del ROPE. El valor de referencia es de 3 cm, un intervalo del 97% y un radio máximo alrededor del ROPE de 7 cm (véase la Fig. 6.123).

Elegimos estos valores arbitrariamente porque nos interesaban ad hoc. Así que no había ninguna razón racional más que demostrar el procedimiento. Los análisis bayesianos permiten calcular probabilidades

adecuadas al caso. Esto significa que no hay una única probabilidad abstracta, sino siempre probabilidades concretas dependientes del contexto.

```
> plotAreaInROPE(meanDiff.h, credMass=0.97, compVal=3, maxROPEradius=7)
```

Otra forma sería contrastar las hipótesis de un modelo entre sí:

```
> # calculate posterior (Odds) Ratios
> p1 <- mean(meanDiff.h > 0)
> p2 <- mean(meanDiff.h < 3)
> p1
[1] 0.7308254
> p2
[1] 0.9371613
> p1vsp2 <- post.comp(post1=p1 , post2=p2, RETURN=TRUE)
Comparison of two posterior values
post[1] = 0.731
post[2] = 0.937
Ratio (p1 vs. p2) = 0.78
Ratio (p2 vs. p1) = 1.28
Odds Ratio (p1 vs. p2) = 0.182
Odds Ratio (p2 vs. p1) = 5.49
```

Pasemos a la implementación del artículo de Bretthorst (1993) en derivación del Mathematica™ de Studer (1998) con las funciones de R `DiffinMeans()` y `UMSprint()`. Además de los datos brutos, elegimos límites superiores e inferiores para la media y la desviación estándar.

```
# Bretthorst 1993
# implementation Mathematica by UM Studer 1999
# prepare data
W.cm <- pres.h.nona[,"W.cm"]
O.cm <- pres.h.nona[,"O.cm"]
Ni <- length(W.cm)
Nii <- length(O.cm)
Di <- mean(W.cm, na.rm=TRUE)
Dii <- mean(O.cm, na.rm=TRUE)
si <- sd(W.cm, na.rm=TRUE)
sii <- sd(O.cm, na.rm=TRUE)
L <- 170
H <- 190
sL <- 4
sH <- 10
```

Les pasamos los valores:

```
inval.base <- list(
Si = NULL, # UMS specific -> successes group1
Ni = Ni, # N group1
Sii = NULL, # UMS specific -> successes group2
Nii = Nii, # N group2
smin = 0, # UMS specific -> bounds on the mean
# only for SucRatesIntBounds()
Di = Di, # mean group1
si = si, # sd group1
Dii = Dii, # mean group2
sii = sii, # sd group2
L = L, # mean lower bound
```

```
H = H, # mean upper bound
sL = sL, # variance lower bound
sH = sH, # variance upper bound
snames = c("president","main opponent")
)
inval.base
```

Y déjalo correr:

```
> DiM.presheights <- DiffinMeans(inval=inval.base, out=FALSE)
L - Mean_comb < 0 : TRUE [comparison L < DD]
'+'-sign between Gamma-factors is ok
Calculate PMV
Calculate PMbarV
Calculate PMVbar
Calculate PMbarVbar
Compile results
Create dataframes for output
```

`UMSprint()` muestra los resultados:

```
> UMSprint(DiM.presheights)
########################################################################
###
### ON THE DIFFERENCE IN MEANS
### G.L. Bretthorst (1993)
###
### original Mathematica code by U.M. Studer (90s, Switzerland)
Note:
If any probability is printed as '1' (= one) or '0' (= zero),
it means that the probability is practically that value by
giving respect to limited computer precision.
-------------------------------- Data (Input) -------------------------
N_1 = 59 : Mean_1 ± SD_1 = 180.30509 ± 7.70637
N_2 = 59 : Mean_2 ± SD_2 = 179.52542 ± 7.01318
N_total = N_1 + N_2 = 118 : Mean_comb ± SD_comb = 179.91525 ± 7.34681
Bounds on the Mean (s_min = 0): Mean_L = 170, Mean_H = 190
Bounds on the Standard Deviation: SD_L = 4, SD_H = 10
Mean_L - Mean_comb < 0 = TRUE (-> '+'-sign between Gamma-fcts o.k.)
-------------------------------- Results ------------------------------
p(mv | D_1, D_2, I) = const. 6.89343e-129
p(mbarv | D_1, D_2, I) = const. 1.3871e-129
p(mvbar | D_1, D_2, I) = const. 3.18009e-129
p(mbarvbar | D_1, D_2, I) = const. 6.43148e-130
where const. = 1.10221e-49 / p(D_1,D_2|I)
= 8.26189e+127
---------- Model ------------------------------ Probability ----------
mv: Same Mean, Same Variance: 0.569528
mbarv: Different Mean, Same Variance: 0.114601
mvbar: Same Mean, Different Variance: 0.262735
mbarvbar: Different Mean, Different Variance: 0.0531361
------------------------------ Odds Ratios ---------------------------
The probability the means are the same is: 0.832263
The probability the means are different is: 0.167737
The odds ratio is 4.962 to 1 in favor of the same means.
The probability the variances are the same is: 0.684129
The probability the variances are different is: 0.315871
The odds ratio is 2.165 to 1 in favor of the same variances
The probability the data sets are the same is: 0.569528
The probability the data sets are different is: 0.430472
The odds ratio is 1.323 to 1 in favor of the same means and variances.
-------------------------------- End ---------------------------------
########################################################################
```

Resulta que un cambio en la Prior puede ir acompañado de un cambio sustancial en la Posterior. Podríamos estrechar o ensanchar los límites a priori y repetir la ejecución. Dejamos eso y la interpretación a los lectores:

```
# different bounds -> smaller
inval.smaller <- inval.base
inval.smaller$L <- 175
inval.smaller$H <- 185
inval.smaller$sL <- 5
inval.smaller$sH <- 9
inval.smaller
DiM.presheights.smaller <- DiffinMeans(inval=inval.smaller, out=FALSE)
UMSprint(DiM.presheights.smaller)
# different bounds -> wider
inval.wider <- inval.base
inval.wider$L <- 160
inval.wider$H <- 200
inval.wider$sL <- 3
inval.wider$sH <- 10
inval.wider
DiM.presheights.wider <- DiffinMeans(inval=inval.wider, out=FALSE)
UMSprint(DiM.presheights.wider)
```

Alternativamente, las mismas funciones R existen en la implementación de los scripts de Gregory's Mathematica™ (2006) llamadas `DiM.pg()`, `PGprint()` y `plot.DiM()`. Estas funciones toman los valores como invtyp = "pg" o como invtyp = "ums", es decir, una vez en el formato original según Gregory y otra vez en el formato utilizado por Studer, tal y como les utiliza la función R `DiffinMeans()` antes mencionada. La salida de `DiM.pg()` se puede representar gráficamente con `plot.DiM()` y corresponde a los gráficos en los gráficos de Bretthorst (1993). Para completar la salida aquí – redundante – también (véase la Fig. 6.124). Los resultados son lógicamente idénticos a los de la implementación de Studer.

```
# implementation Mathematica by Phil Gregory R-Code
inputvalues <- list(
snames = c("president","opponent"),
# sample 1
d1 = W.cm[!is.na(W.cm)],
# sample 2
d2 = O.cm[!is.na(O.cm)],
# Input priors and no. of steps in evaluation of
# p(r|D_1,D_2,I) & p(delta|D_1,D_2,I)
# ndelta = number of steps in delta parameter
# (mean difference)
ndelta = 1000, #100
# nr = number of steps in r parameter
# (ratio of the standard deviations)
nr = 1000, # 100
# Set prior limits (assumed the same for each data set)
# on mean (low,high),
# and prior limits (assumed the same for each data set)
# on the standard deviation (sigmalow, sigmahigh).
# upper mean
high = 200,
# lower mean
low = 160,
# upper sd
sigma.high = 10,
# lower sd
sigma.low = 3
)
# according to Phil Gregory scheme
```

```
inputvalues
DiM.PG.presheights <- DiM.pg(invtyp="pg",
inputvalues=inputvalues,
print.res=TRUE)
DiM.print.pg(DiM.PG.presheights)
# determine low + high from means
# determine scaleL and scaleH by experimenting
```

Podemos trazarlo ajustando primero el gráfico a los valores que nos resulten importantes:

```
> # check before
> # change = TRUE -> change to new values
> DiM.presheights.newlimits <- DiM.extract.limits(
        + DiM.PG.presheights,
        + low=175, high=182,
        + scaleL=2, scaleH=2,
        + change=TRUE)
low = 175
high = 182
ndelta = 1000
sigma.low = 3
sigma.high = 10
delta.low = -7
delta.high = 7
delta.delta = 0.01401
delta.sek.l = 1000
scaleL = 2
scaleH = 2
d1std = 7.706
d2std = 7.013
r.low = 0.5494
r.high = 2.198
r.delta = 0.00165
r.sek.l = 1000
fac.brob = 1
> DiM.presheights.plotvalues <- DiM.plot.calc.pg(
        + DiM.presheights.newlimits,
        + BROB=FALSE)
Calculate graphical output for the various probabilities...
Calculate pdel.SD1D2I and pdelA...
Calculate pdel.SbarD1D2I and pdelB...
Calculate average of pdelA and pdelB...
Calculate HDIs for the average of pdelA and pdelB...
Calculate pr.CD1D2I and prA...
Calculate pr.CbarD1D2I and prB...
Calculate average of prA and prB...
Calculate HDIs for the average of prA and prB...
Compile results...
```

Y después trazarlo:

```
R-Code DiM.plot.pg(DiM.presheights.plotvalues,
        filling=FALSE, BROB=FALSE)
```

O trazarlo directamente (véase Fig. 6.124):

```
DiM.presheights.plotvalues <- DiM.plot.calc.pg(DiM.PG.presheights,
        low=175, high=182, scaleL=2, scaleH=2, BROB=FALSE)
DiM.plot.pg(DiM.presheights.plotvalues, filling=FALSE, BROB=FALSE)
```

Las figuras 6.124 y 6.125 muestran las seis hipótesis del problema de Behrens-Fisher (Bretthorst, 1993). La figura 6.124 contiene las hipótesis sobre las diferencias medias, mientras que la figura 6.125 visualiza las relativas a las desviaciones típicas:

- Diferencias medias bajo el supuesto de que las desviaciones estándar son iguales (arriba).
- Diferencias medias suponiendo que las desviaciones estándar son diferentes (centro)
- Diferencias medias independientemente de que las desviaciones estándar sean iguales o diferentes (abajo)

y

- Desviaciones estándar suponiendo que las diferencias medias son iguales (arriba)
- Desviaciones estándar suponiendo que las diferencias medias son diferentes (centro)
- Desviaciones estándar sin tener en cuenta si las diferencias medias son iguales o diferentes (abajo). son diferentes (abajo)

Debe hacerse una advertencia a ambas funciones de R `DiM.pg()` y `DiffinMeans()`: Dado que las funciones internas de T se utilizan exageradamente para formar las integrales respectivas, los scripts de R desgraciadamente salen con un error, ya que el límite de computabilidad del ordenador se sobrepasa fácilmente en la dirección del infinito. Este límite se encuentra en

```
# infinity is (not) the end...
# use BROB
# for very large numbers
print(10^(1.310))
print(as.brob(10)^(1.3100))
```

es decir, 10e + 308 y a partir de 10e + 309 la R sigue calculando en el infinito. Dada la suposición de que el número de átomos en el universo conocido se estima entre 10e + 84 y 10e + 89 con cierta imprecisión, esto ya es un número decentemente grande. En estos casos, nos referimos a `brm()` del paquete de R `brms` o BUGS, JAGS o el paquete de R `BEST`, que son todas implementaciones MCMC que no conocen o eluden este problema realizando cálculos en el plano logarítmico.

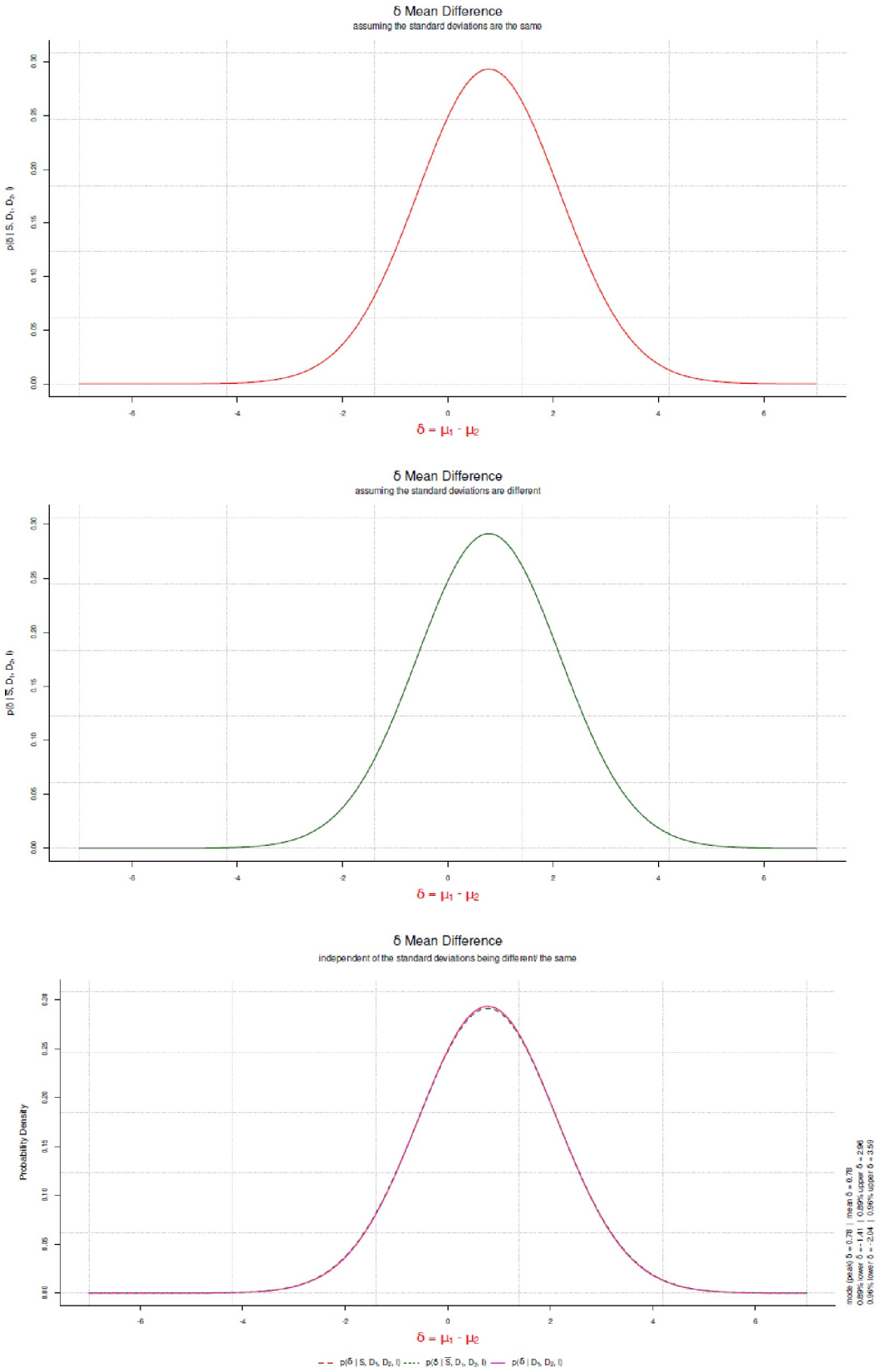


Figura 6.124. *Presidentes de EE.UU. y aspirantes*
(alturas corporales, prueba t bayesiana según Bretthorst 1993, Posteriores según hipótesis)

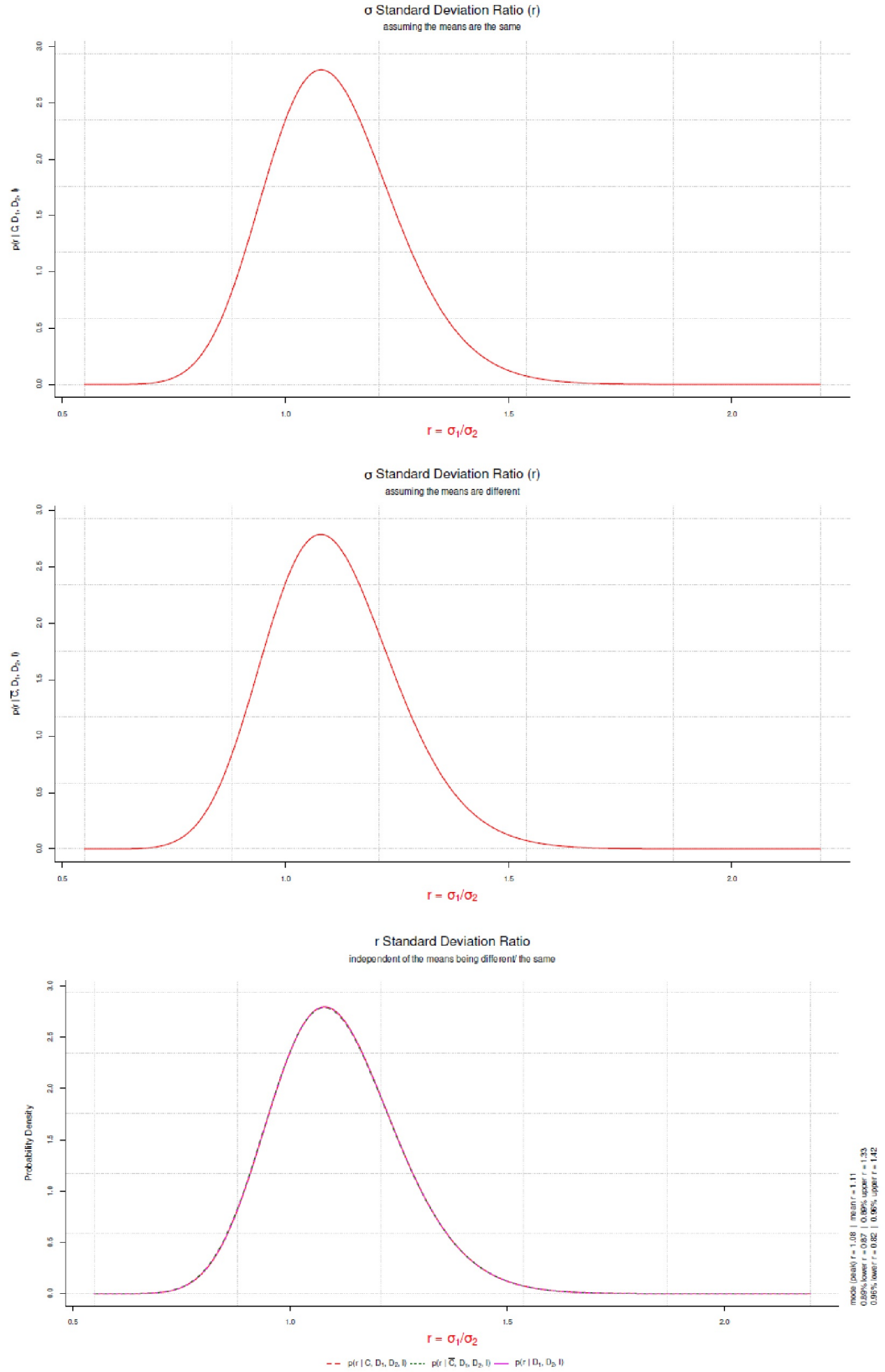


Figura 6.125. *Presidentes de EE.UU. y aspirantes*
(alturas corporales, prueba t bayesiana según Bretthorst 1993, Posteriores según hipótesis)

Como conclusión a este estudio de caso, se pueden examinar los individuos realizando un bootstrap bayesiano (Rubin, 1981). Se utiliza la implementación `bayesboot()` del paquete `bayesboot` de R. Se realiza un bootstrap por separado para *W.red* y *O.red* utilizando los valores medios reponderados. Podemos volver a examinar los casos,

```
# cases
pres.Bboot.W <- bayesboot(W.cm[!is.na(W.cm)], mean)
pres.Bboot.O <- bayesboot(O.cm[!is.na(O.cm)], mean)
```

pero en lo que sigue nos basamos en el análisis de las personas:

```
> # persons
> pres.Bboot.W <- bayesboot(dats$W.red, mean)
> pres.Bboot.O <- bayesboot(dats$O.red, mean)
> summary(pres.Bboot.W)
Bayesian bootstrap
Number of posterior draws: 4000
Summary of the posterior (with 95% Highest Density Intervals):
statistic mean     sd       hdi.low  hdi.high
V1        179.6597 1.201954 177.3905 182.0223
Quantiles:
statistic q2.5%  q25%    median   q75%     q97.5%
V1        177.29 178.873 179.6732 180.4481 181.9605
Call:
bayesboot(data = dats$W.red, statistic = mean)
> summary(pres.Bboot.O)
Bayesian bootstrap
Number of posterior draws: 4000
Summary of the posterior (with 95% Highest Density Intervals):
statistic mean     sd       hdi.low  hdi.high
V1        179.5326 1.002259 177.7032 181.5263
Quantiles:
statistic q2.5%    q25%     median   q75%     q97.5%
V1        177.5875 178.8323 179.5204 180.2014 181.468
Call:
bayesboot(data = dats$O.red, statistic = mean)
```

Y trazamos:

```
R-Code plot(pres.Bboot.W)
plot(pres.Bboot.O)
```

También se podría conseguirlo mediante valores iniciales ponderados

```
# boot via re-weighting original data set
pres.Bboot.W.1 <- bayesboot(dats$W.red, weighted.mean,
        use.weights=TRUE)
pres.Bboot.O.1 <- bayesboot(dats$O.red, weighted.mean,
        use.weights=TRUE)
```

Las dos distribuciones posteriores pueden restarse entre sí para formar un índice de comparación.

```
# compare
difference <- pres.Bboot.W.1-pres.Bboot.O.1
pres.Bboot.diff <- as.bayesboot(difference)
```

El resultado se puede representar gráficamente, se pueden obtener estadísticas resumidas y se pueden calcular los HDI.

```
> summary(pres.Bboot.diff)
Bayesian bootstrap
Number of posterior draws: 4000
Summary of the posterior (with 95% Highest Density Intervals):
statistic   mean      sd        hdi.low  hdi.high
V1          0.1078175 1.574035 -2.941893 3.146458
Quantiles:
statistic q2.5%    q25%      median    q75%    q97.5%
V1       -2.92777 -0.9243786 0.1454041 1.17227 3.174135
Call:
> attr(pres.Bboot.diff,"statistic.label") <- "Difference"
```

La Fig. 6.126 muestra el gráfico usual:

```
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l")
plot(pres.Bboot.diff)
```

Del mismo modo, se podrían probar hipótesis sobre las distribuciones, como si normalmente los ganadores o los retadores son más grandes:

```
> prob.winner.is.larger <- mean(pres.Bboot.W.1[,1] > Output
+ pres.Bboot.O.1[,1])
> prob.winner.is.larger
[1] 0.5325
> # OR
> prob.winner.is.larger / (1-prob.winner.is.larger)
[1] 1.139037
> 1 / (prob.winner.is.larger / (1-prob.winner.is.larger))
[1] 0.8779343
```

O mayor que un varón medio en EE.UU. (Frya, Gu & Ogden, 2012):

```
> # Given the model and the data,
> # this is the probability that the mean
> # heights of American presidents is above the mean heights of
> mean(pres.Bboot.W[,1])
[1] 179.6597
> mean(pres.Bboot.O[,1])
[1] 179.5326
> mean(c(pres.Bboot.W[,1] > 175.9, TRUE, FALSE))
[1] 0.9970015
> mean(c(pres.Bboot.O[,1] > 175.9, TRUE, FALSE))
[1] 0.9995002
```

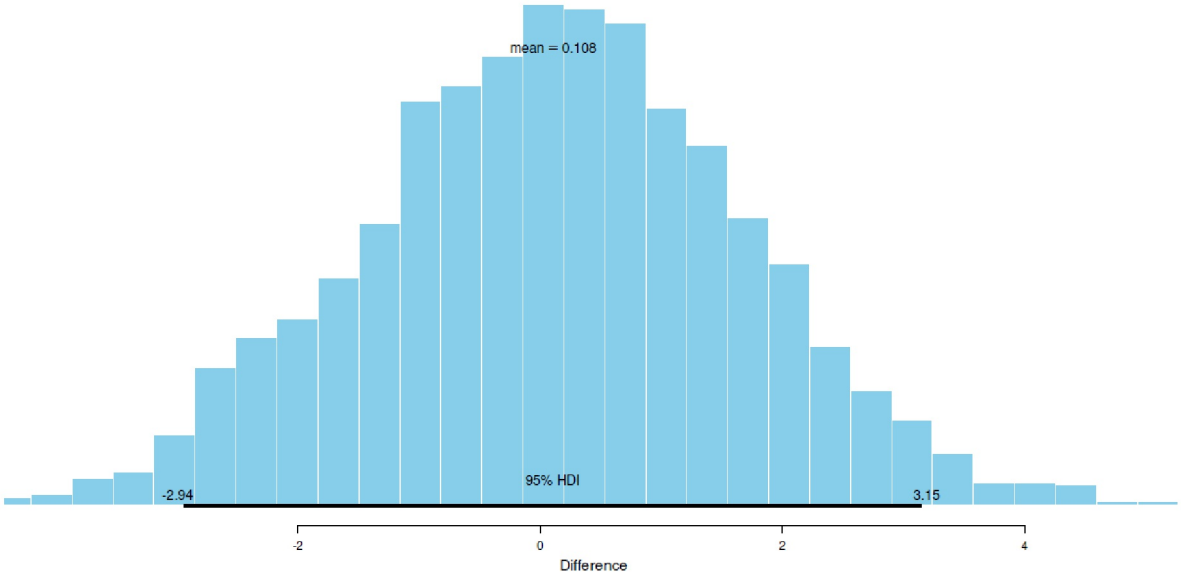


Figura 6.126. Presidentes de EE.UU. y aspirantes
(tamaños corporales, bootstrap bayesiano, diferencias posteriores)

En comparación, se ve diferente cuando la población mundial sirve como sistema de referencia

```
> # US
> p1 <- mean(pres.Bboot.W[,1] > 176.4) #2011-2014
> p2 <- mean(pres.Bboot.O[,1] > 176.4) #2011-2014
> p1
[1] 0.994
> p2
[1] 0.9995
> p1vsp2 <- post.comp(post1=p1 , post2=p2, RETURN=TRUE)
Comparison of two posterior values
post[1] = 0.994
post[2] = 1
Ratio (p1 vs. p2) = 0.994
Ratio (p2 vs. p1) = 1.01
Odds Ratio (p1 vs. p2) = 0.0829
Odds Ratio (p2 vs. p1) = 12.1
```

o los habitantes de Suecia (Wikipedia, 2019h):

```
> # Sweden
> p1 <- mean(pres.Bboot.W[,1] > 181.5)
> p2 <- mean(pres.Bboot.O[,1] > 181.5)
> p1
[1] 0.0625
> p2
[1] 0.023
> p1vsp2 <- post.comp(post1=p1 , post2=p2, RETURN=TRUE)
Comparison of two posterior values
post[1] = 0.0625
post[2] = 0.023
Ratio (p1 vs. p2) = 2.72
Ratio (p2 vs. p1) = 0.368
Odds Ratio (p1 vs. p2) = 2.83
Odds Ratio (p2 vs. p1) = 0.353
```

Para la comparación de ganador vs. retador, se puede hacer un gráfico para diferentes diferencias, que muestra la diferencia en cm en el eje X y la probabilidad de que la masa de la Posterior esté por encima de esta diferencia en el eje Y (ver Fig. 6.127, R-código no impreso). De este modo, examinamos la probabilidad con la que el ganador (o perdedor) es mayor.

Recordamos que ahora se trata de los dos grupos en general (= personas) y no de enfrentamientos por año electoral. La curva roja muestra la probabilidad de que los ganadores (independientemente de su afiliación partidista) sean mayores que los contrincantes. La curva azul muestra lo contrario, es decir, cuando los retadores son mayores que los ganadores. Por lo tanto, las dos curvas no son más que imágenes especulares la una de la otra. Algunos valores de ejemplo: con una probabilidad del 53.25%, los ganadores son al menos tan grandes o más que los aspirantes. Con un 11.2%, los ganadores son al menos 2 cm más altos que los aspirantes y con un 9.9% son al menos 2 cm más bajos. Como en todas las demás simulaciones MCMC, recordemos que se trata de simulaciones, no de resultados analíticos exactos. Es decir, las repeticiones por parte de los lectores darán lugar aquí a pequeñas desviaciones, como muestran las siguientes repeticiones de los cálculos completos. Primero una ejecución con los resultados que se acaban de comunicar

```
> # prob that mean diff is above 2 cm
> mean(pres.Bboot.diff[,1] >= 0)
[1] 0.5325
> mean(pres.Bboot.diff[,1] >= 2)
[1] 0.112
> mean(pres.Bboot.diff[,1] <= -2)
[1] 0.099
```

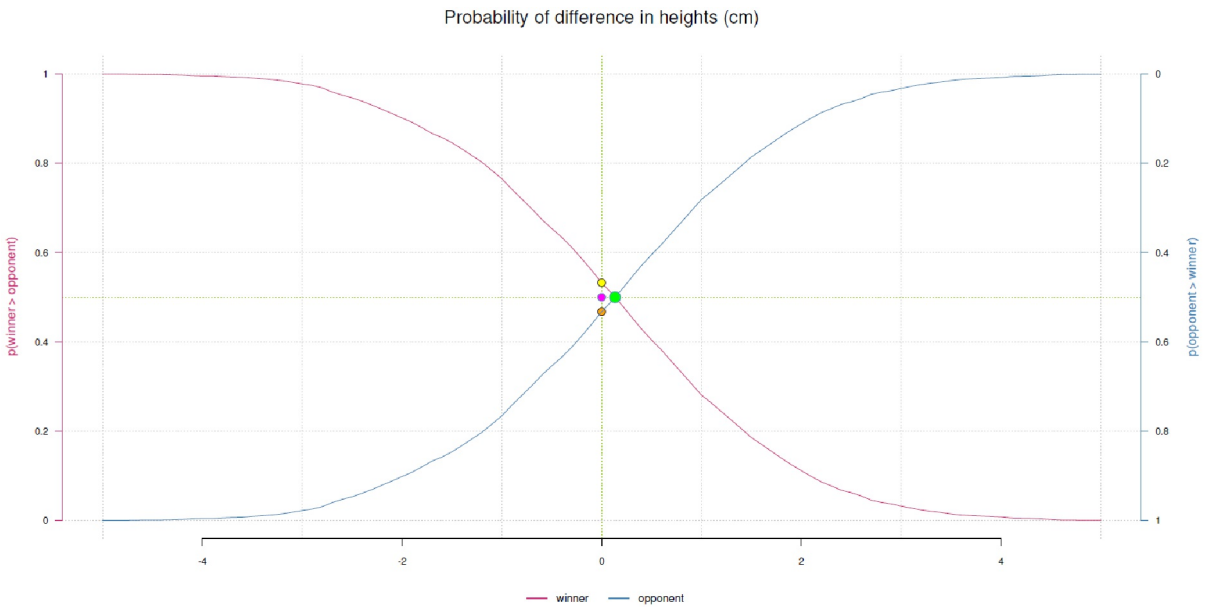


Figura 6.127. *Presidentes de EE.UU. y aspirantes*
(alturas corporales, bayesian bootstrap, probabilidades de las diferencias)

y luego repetimos otra vez:

```
> # prob that mean diff is above 2 cm
> mean(pres.Bboot.diff[,1] >= 0)
[1] 0.52075
```

```
> mean(pres.Bboot.diff[,1] >= 2)
[1] 0.1115
> mean(pres.Bboot.diff[,1] <= -2)
[1] 0.08625
```

Los resultados son similares, pero no idénticos. No se trata de una inexactitud en sentido estricto, sino que forma parte del sistema de simulación de datos. Las diferencias en las estimaciones son

```
> # differences, secondary run, just example results
> 0.5325-0.52075 #>=0
[1] 0.01175
> 0.112-0.1115 #>=2
[1] 5e-04
> 0.099-0.08625 #<=-2
[1] 0.01275
```

En comparación, las desviaciones estándar de las hipótesis son significativamente mayores

```
> # sd that mean diff is above 2 cm
> sd(pres.Bboot.diff[,1] >= 0)
[1] 0.4996317
> sd(pres.Bboot.diff[,1] >= 2)
[1] 0.3147896
> sd(pres.Bboot.diff[,1] <= -2)
[1] 0.280768
```

de modo que aceptamos estas variaciones de MCMC a MCMC como parte de la incertidumbre cotidiana. Aquí vemos de nuevo si de estas diferencias surge o no información con significado. En nuestro caso, consideramos que las fluctuaciones carecen de significado.

Del mismo modo, puede merecer la pena explorar más de cerca la distinción entre personas y elecciones en términos de las hipótesis ya mencionadas. Una comparación apunta a un efecto global en todas las personas, la otra examina cada encuentro electoral individualmente.

Otro tipo de representación (véase la Fig. 6.128):

```
# height ratio R-Code
# http://www.nicebread.de/a-compendium-of-clean-graphs-in-r/
# Presidential data up to and including 2008
# data from Stulp et al. 2013
# cases ie. elections
pres.h.nona$W.O.ratio <- pres.h.nona$W.cm / pres.h.nona$O.cm
pres.h.nona$W.O.ratio.TF <- pres.h.nona$W.O.ratio > 1
head(pres.h.nona)
colos <- ifelse(pres.h.nona$W.O.ratio.TF, "blue","red")
# pres.h.nona.ratio = ratios without NAs
# get rid of NAs
pres.h.ratio.nona <-
       pres.h.nona$W.O.ratio[!is.na(pres.h.nona$W.O.ratio)]
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l", mfrow=c(1,2))
plotPost(pres.h.ratio.nona, xlab="ratio",
       ylab="Density", cex.lab=1.2)
lines(density(pres.h.ratio.nona), col="darkred", lty=1, lwd=2)
with(pres.h.nona, plot(year, W.O.ratio, pch=21, col=colos,
       bg=colos, pre.plot=grid(), bty="n",
       xlab="year", ylab="Winner/ Main Opponent",
       type="b", cex.lab=1.2) )
abline(h=1, col="grey", lty=2)
mtext("Presidential heights (ratios)", outer=TRUE,
       line=-2, cex=1.5, side=3)
```

Aún podemos investigar qué tipo de secuencias ha habido a lo largo de los años. Aquí examinamos con qué frecuencia se produjo un cambio, es decir, con qué frecuencia en las siguientes elecciones el siguiente presidente volvió a ser mayor que el aspirante (el código R necesario para generar los valores de verdad no se imprime (véase `ptII_quan_Bayes_case_presidential-heights.r`).

```
> TF.no.chars.W.tab
TF.no.chars.W
 0 1 2 3 5 7
15 6 2 3 1 1
> TF.no.chars.0.tab
TF.no.chars.0
 0 1 2 3 5
19 6 3 2 2
```

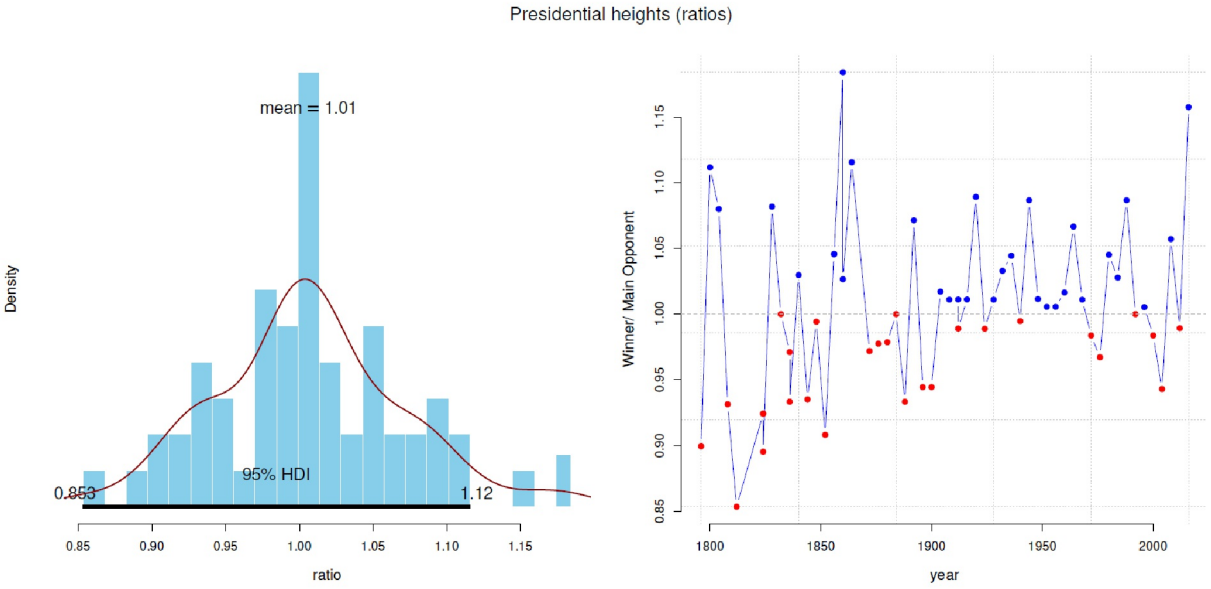


Figura 6.128. *Presidentes de EE.UU. y aspirantes*
(alturas corporales, bootstrap bayesiano, proporciones por elección)

Los resultados muestran que existen variantes de secuencia de 1, 2, 3, 5 y 7 elecciones consecutivas en las que el presidente fue consecutivamente mayor que el aspirante. A la inversa, hay variantes de secuencia de 1, 2, 3 y 5 elecciones consecutivas en las que el aspirante fue mayor que el presidente electo. Si se suma el número de variantes de secuencia de cada una, se obtiene el mismo número para ambos casos, sólo que distribuido de forma diferente.

```
> # without zero
> sum(TF.no.chars.W.tab[-c(1)])
[1] 13
> sum(TF.no.chars.0.tab[-c(1)])
[1] 13
```

Lo más interesante es el significado que se esconde tras ese análisis, que dejamos a los lectores. Por supuesto, se podrían examinar los contextos desde el punto de vista sociológico, político, económico, etc., para ver si se puede formar aquí una hipótesis significativa. Lo dudamos.

Tarea 6.12: Convertirse en POTUS

Las explicaciones anteriores, junto con otros datos (por ejemplo, sobre el comportamiento electoral), pueden conducir a nuevas preguntas: cómo se relaciona la proporción de la altura corporal de las personas que se reúnen en cada caso con el número de votos, y mucho más. Es significativo que el número de votos electorales no significa que se haya ganado la elección, ya que la elección del presidente de EE.UU. se realiza a través del llamado colegio electoral y no no refleja una función directa de la mayoría de los votos escrutados.

Las razones residen en la ley electoral estadounidense. Se requieren 270 votos electorales, cuyo número varía según el estado (Oelker & Klormann, 2012-11-06, para más detalles). La ley electoral estadounidense permite que una persona se convierta en presidente de EE.UU. aunque reúna menos votos en términos absolutos que su contrincante. Así ocurrió, por ejemplo, en las elecciones de 2018 entre Clinton y Trump, en las que Trump y los republicanos obtuvieron menos votantes pero más votos electorales. La clave está en la forma en que se trazan los distritos electorales, que los partidos pueden utilizar sistemáticamente para manipularlos. Los distritos se construyen de tal manera que se agrupan muchos votantes del partido contrario, pero luego reciben pocos votos electorales. Y a la inversa, los distritos se crean de tal manera que unos pocos votantes del propio partido puedan contar tantos votos electorales como sea posible. Los lectores interesados pueden examinar esta relación con más detalle.

6.15.2 Estudio de caso: índices de aprobados en el tratamiento de drogodependencias

Este estudio de caso examina cómo puede fluir información cualitativa diferente en una distribución a priori y cómo es la distribución posterior resultante. A continuación se amplía el procedimiento descrito por Urban Studer. Se incluyen los valores empíricos de la época del estudio de Studer (1998), así como datos más recientes de casi dos décadas, que la institución, el centro suizo de tratamiento hospitalario de la drogadicción `start again`, tuvo la amabilidad de poner a nuestra disposición. A través del aspecto a largo plazo, se puede separar bien el desarrollo y la separación de las fluctuaciones anuales de las tendencias a largo plazo. Studer (1998), como físico matemático, utiliza el enfoque del principio de máxima entropía para formular la Prior. En otro documento de trabajo (1996b) describe los pasos cualitativos previos a la aplicación de la máxima entropía para acotar la Prior para distintos escenarios (por ejemplo, el éxito y el fracaso son posibles en principio o no está claro si el éxito y el fracaso son posibles en principio).

Studer (1996b, 1998) llevó a cabo una evaluación a largo plazo de `start again` con métodos mixtos entre 1992 y 1998. El estudio fue financiado por la Oficina Federal de Justicia (BAJ) para evaluar el éxito del enfoque *terapia en lugar del castigo* para los clientes drogodependientes. Entre otras cosas, el estudio abordaba cuestiones estadísticas de *éxito básico* o *las diferencias entre los subgrupos de clientes* (por ejemplo, clientes de medidas frente a clientes voluntarios, hombres frente a mujeres, éxito de la terapia del enfoque sistémico profundo en sí, ...). No profundizamos en las diferencias entre subclientes. Para más detalles, véase Studer (1998). El *éxito* como criterio se operacionalizó de estas maneras:

- **Tasa de finalización terapéutica** – Si un cliente completó todo el programa terapéutico o lo abandonó. Studer (1998) distingue entre abandono temprano y tardío.
- **Periodo catamnésico** – En el que los antiguos clientes fueron examinados catamnésicamente tras un periodo medio de 1.66 ± 0.44 años (abandonos tempranos, $n = 58$) o 1.93 ± 0.50 años (abandonos tardíos, $n = 23$). Junto con otros abandonos que se incluyeron poco antes del final del estudio ($n = 8$, periodo de catamnesis 0.50 ± 0.91 años), resulta una muestra total de 89 antiguos clientes.

- **Éxito catamnésico** – Los clientes se clasificaron en función de si podían asignarse a las categorías *fracaso, indiferente* o *éxito* (Studer, 1998, capítulo 10.1 para las definiciones exactas de los criterios de éxito formulados de forma conservadora).

En un informe puramente metodológico (Studer, 1996b), Studer se adentra en consideraciones preliminares bayesianas para explicar los conocimientos previos sobre el éxito terapéutico en la terapia de la adicción mediante la Máxima entropía. Las afirmaciones se concentran inicialmente en la tasa de rendimiento terapéutico. Las explicaciones son transferibles al éxito catamnésico, ya que se aplica la misma lógica. Estos son los argumentos de Studer:

- El éxito terapéutico es posible en principio en casos individuales. Pero es igualmente posible que este proceso fracase. Por tanto, el fracaso también es posible en principio.
- Tanto el fracaso como el éxito son plausibles y se han observado e informado repetidamente en una amplia variedad de estudios sobre adicción y en la práctica clínica. Ambos resultados se consideran posibles. Sería aún más inespecífico si no hiciéramos esta suposición y asumiéramos que ni siquiera sabemos si el éxito o el fracaso son posibles en la terapia de la adicción. Studer (1996b) ofrece fórmulas para ambos casos.
- Como resultado, los valores entre 0 y 1 pueden asumirse como un rango válido de valores de éxito terapéutico. Tanto el éxito como el fracaso pueden expresarse directamente como probabilidad. Si no sabemos si es posible el éxito o el fracaso, pero excluimos ambos valores del continuo de valores válidos, la posterioridad no consideraría estos valores más allá. Si entonces nos equivocamos, los resultados estarán masivamente sesgados. Por lo tanto, reducir demasiado la Prior requiere buenas razones. Esto no se opone a una estimación conservadora de la Prior, por ejemplo en lo que se refiere al grueso de su masa.
- Sin embargo, tras el primer fracaso o los primeros fracasos, no debería estallar la depresión en la institución de que el concepto no funciona. Del mismo modo, tras el/los primer(os) éxito(s), no debería estallar la euforia de que todos los clientes siempre tienen éxito. Esto puede servir para reducir aún más el abanico general de valores: En un entorno terapéutico serio no cabe esperar sólo éxitos o sólo fracasos. El valor real se encuentra entre ambos, es decir, entre 0 y 1.
- Asimismo, no existen clientes idénticos, es decir, los procesos interindividuales y el resultado final varían y el tiempo en la institución también varía. El éxito depende del tiempo, al igual que el fracaso. En sentido estricto, sólo se puede evaluar retrospectivamente el éxito cuando se observa la muerte (y entonces por drogas o no). Estrictamente hablando, los éxitos (catamnésicamente) tienen que estar marcados en el tiempo y contabilizados con muchos otros factores influyentes. En este caso, sólo se trata de índices de éxito con un marco temporal relativamente manejable, la duración de la estancia en el centro. Estas ampliaciones (necesarias para las catamnesis) no son necesarias para esto. En Studer (1996b, 1998) se encuentra más detalles sobre cómo resuelve el autor este problema con respecto a la tasa de éxito terapéutico en el intervalo de tiempo de catamnesis definido en su obra (véanse también las explicaciones en el Cap. 5.5.7 sobre el AED y el éxito catamnésico). Esto no es importante para las explicaciones siguientes.

Sin embargo, lo que parece trivial y podría surgir del sentido común se ha puesto ahora en una forma científicamente accesible y bien fundamentada: la tarea realizada. A partir de ahí se pueden formular derivaciones. Studer (1996b, p.26f.) elabora matemáticamente, según dos principios, la elección de dos Priores diferentes en los que desembocan las consideraciones anteriores y que representan una ignorancia completa:

- si no hay razón para preferir un rango de valores, pero el éxito y el fracaso parecen posibles, se puede aplicar el principio de razón insuficiente de Bayes (1763), Bernoulli (1713) y Laplace (1843), respectivamente;
- si no está claro si es posible el éxito o el fracaso, se utiliza una distribución previa según Jeffreys (1939/1961) y Carnap (1952).

Ambas variantes contienen un mínimo de información específica. Tomamos las fórmulas para calcular las tasas de éxito (= proceso binomial) de Studer (1996b, cap. 5) y las implementamos en el programa R. Si no sabemos en absoluto si es posible el éxito o el fracaso, resulta la siguiente distribución "pre-prior" con

probabilidad de éxito p_{JC}. *H* denota la hipótesis sobre el modelo bayesiano e *I* la información disponible, es decir, el conocimiento pevio. θ representa el parámetro desconocido del modelo *H*

$$p_{\mathrm{JC}}(H \mid I) = \frac{\text{const.}}{\theta \cdot (1-\theta)}, \quad \text{für } 0 < \theta < 1 \tag{6.172}$$

Studer lo denomina Bayes-Laplace. El valor p_{JC} expresa la ignorancia completa con respecto al parámetro. Si ahora establecemos la restricción de que tanto el éxito como el fracaso son siempre posible, que corresponde al *principio de razón insuficiente* de Bernoulli, obtenemos la siguiente distribución a priori de la probabilidad de éxito p_{BL}:

$$p_{\mathrm{BL}}(H \mid I) = 1, \quad \text{für } 0 \leq \theta \leq 1 \tag{6.173}$$

Studer la denomina Jereys-Jaynes. El valor p_{BL} corresponde a una distribución uniforme. Según el teorema de Bayes, los datos empíricos de éxito o fracaso se pueden utilizar ahora para calcular las probabilidades posteriores de los dos casos. Dado que existe una solución analítica al teorema de Bayes para este caso binomial, no es necesario utilizar simulaciones MCMC, que por supuesto serían posibles sin más. Para el caso de probabilidad – binomial – y cuando *s = aciertos* y *N = número total de casos* recurrimos a

$$p(s/N \mid H, I) = \frac{N!}{s! \cdot (N-s)!} \cdot \theta^s \cdot (1-\theta)^{N-s}, \quad \text{für } s = 0, 1, \ldots, N \tag{6.174}$$

Esto corresponde al proceso Bernoulli con los pesos θ o (*1*- θ). El proceso describe una secuencia dicotómica de sucesos que toman el valor 1 (= éxito) o el valor 0 (= fracaso) con la probabilidad θ o (*1*-θ). Además, se supone que los sucesos son independientes entre sí y que el proceso es estacionario (= constante) y no cambia. Como subraya el autor, esta distribución de probabilidad puede ser generada por un proceso Bernoulli bien fundado, pero esto podría ser problemático (Studer, 1996b, p.26, cursiva en el original).

> "La cuestión, sin embargo, de si tal proceso subyace realmente a la situación real del examen – aunque se confirmara experimentalmente la distribución (5.11) – es ¡otra diferente! JAYNES[...] ha argumentado a favor de la suposición ciega de que tales procesos aleatorios necesariamente subyacen a una situación dada, acuñó la expresión *Mind Projection Fallacy*: ¡No debemos proyectar irreflexivamente nuestras ideas (modelo) y nuestra ignorancia a la naturaleza para verlas como propiedades reales o como indeterminación/aleatoriedad esencial de la naturaleza!
>
> Por tanto, preferimos la línea de argumentación de la teoría de la información basada en el como instrumento para generar distribuciones de máxima adecuación/equidad a los datos reales disponibles – la línea de pensamiento ortodoxa, basada en procesos aleatorios modelizados".

A continuación, Studer (1996b) analiza varios casos límite de las respectivas distribuciones a priori y sus implicaciones. Las distribuciones a posteriori para determinar las probabilidades de éxito se derivan de

$$p_{\mathrm{JC}}(H \mid s/N, I) = \frac{(N-1)!}{(s-1)! \cdot (N-s-1)!} \cdot \theta^{s-1} \cdot (1-\theta)^{N-s-1}, \quad \text{für } 0 \leq \theta \leq 1 \tag{6.175}$$

o de

$$p_{\mathrm{BL}}(H \mid s/N, I) = \frac{(N+1)!}{s! \cdot (N-s)!} \cdot \theta^s \cdot (1-\theta)^{N-s}, \quad \text{für } 0 \leq \theta \leq 1,\ s \geq 1,\ N-s \geq 1 \tag{6.176}$$

Se puede calcular los valores esperados y las varianzas mediante

$$\bar{\theta}_{\mathrm{JC}} = \frac{s}{N} \tag{6.177}$$

$$\sigma^2_{\mathrm{JC}} = \frac{\bar{\theta}_{\mathrm{JC}} \cdot (1 - \bar{\theta}_{\mathrm{JC}})}{N + 1} \tag{6.178}$$

o por

$$\bar{\theta}_{\mathrm{BL}} = \frac{s + 1}{N + 2} \tag{6.179}$$

$$\sigma^2_{\mathrm{BL}} = \frac{\bar{\theta}_{BL}(1 - \bar{\theta}_{BL})}{N + 3} \tag{6.180}$$

La figura 6.129 muestra ambas distribuciones para el caso $s = 23$ aciertos con $N = 27$ casos. Es posible determinar los HDI. Studer elige el 69%, ya que corresponde a una desviación estándar en cada dirección, así como los típicos 95% y 99%. Se tiene que subrayar sus conclusiones sobre el pensamiento práctico cotidiano. Dado el caso en que no hay éxito (s = 0): mientras que la Prior de Jeffreys-Carnap lleva a $\theta_{(JC,\, s=0)} = 0$ y $\sigma^2_{(JC,\, s=0)} = 0$, la Prior Bayes-Laplace para este caso, sin embargo, arroja

$$\bar{\theta}_{(\mathrm{BL},\, s=0)} = \frac{1}{N + 2} \tag{6.181}$$

$$\sigma^2_{(\mathrm{BL},\, s=0)} = \frac{1}{N + 2} \cdot \sqrt{\frac{N + 1}{N + 3}} \tag{6.182}$$

Por lo tanto, si nos enfrentamos a la situación de que no tenemos éxito – no sólo en el caso de los índices de aprobados en el tratamiento de la drogodependencia – esto no es motivo de depresión, pero no deja de ser una situación grave. Podemos, según lo anterior, calcular un intervalo alrededor de $\theta_{(BL,\, s=0)}$ con la ayuda de $\sigma^{2(BL,}_{\ \ s=0)}$, que incluirá el 0 como valor, pero dejará una pequeña cantidad en la parte superior. De todos modos, si N es un número pequeño, ni el apocalipsis ni la euforia son especialmente acertados, porque se trata de valores robustos (que aún no están disponibles), ya que las condiciones cambian muy deprisa, o a menudo incluso cambian con demasiada rapidez.

Ya se pueden utilizar varias funciones R: `pbl()` y `pjc()` funcionan según el principio de aproximación de rejilla (véase también McElreath, 2015, capítulo 2.4.1). Esto significa que emiten la probabilidad posterior para un valor dado θ según las fórmulas anteriores. Si la función se aplica ahora para todo el intervalo de valores de θ, es decir, $0 < \theta < 1$, la se crea una red de probabilidades posteriores, que es una aproximación relativamente buena a la distribución de la probabilidad posterior total. Con, por ejemplo 1 000 puntos de datos, se está en un rango bastante bueno. Según McElreath (2015), ya existe una buena aproximación a partir de 100 puntos de datos. `plot.bl.jc()` traza las dos funciones una al lado de la otra y `sN.ME.ME()` y `sN.ME.post.summary()` calcula un resumen estadístico para la tasa de éxito s_i y el número de intentos N_i. A continuación reproducimos la tasa de éxito 23 de 27 casos según Studer (1996b). Las funciones nos permiten trabajar con el logaritmo o dar salida al resultado en la escala `log()` o transformarlo hacía atrás mediante `exp()`. Como también se puede ver en la Figura 6.129, los resultados son idénticos en cada caso (`ptII_quan_Bayes_case_startagain-successrates.r`).

```
> steps <- 1000
> theta <- seq(0,1,length.out=steps)
> pbl.res <- pbl(theta=theta, si=si, Ni=Ni, loga=TRUE, reexp=TRUE)
> pjc.res <- pjc(theta=theta, si=si, Ni=Ni, loga=TRUE, reexp=TRUE)
> head(pbl.res)
[1] 0.000000e+00 5.008286e-64 4.184442e-57
[4] 4.677278e-53 3.482083e-50 5.875199e-48
> tail(pbl.res)
[1] 2.747476e-04 1.151696e-04 3.729209e-05
```

```
[4] 7.538338e-06 4.821358e-07 0.000000e+00
> head(pjc.res)
[1] 0.000000e+00 6.094746e-62 2.548645e-55
[4] 1.901120e-51 1.062559e-48 1.435698e-46
> tail(pjc.res)
[1] 6.713894e-03 3.514404e-03 1.515770e-03
[4] 4.591425e-04 5.867267e-05 0.000000e+00
> sN.ME.res <- data.frame(pbl.res, pjc.res)
> head(sN.ME.res)
  pbl.res      pjc.res
1 0.000000e+00 0.000000e+00
2 5.008286e-64 6.094746e-62
3 4.184442e-57 2.548645e-55
4 4.677278e-53 1.901120e-51
5 3.482083e-50 1.062559e-48
6 5.875199e-48 1.435698e-46
> tail(sN.ME.res)
     pbl.res      pjc.res
995  2.747476e-04 6.713894e-03
996  1.151696e-04 3.514404e-03
997  3.729209e-05 1.515770e-03
998  7.538338e-06 4.591425e-04
999  4.821358e-07 5.867267e-05
1000 0.000000e+00 0.000000e+00
> plot.bl.jc(theta, sN.ME.res=sN.ME.res, si=si, Ni=Ni, filling=FALSE)
> plot.bl.jc(theta, sN.ME.res=sN.ME.res, si=si, Ni=Ni, filling=TRUE)
> sN.ME.post.summary <- sN.post.su(Ni=Ni, si=si)
################################
OUTPUT posterior results and HDI
Bayes-Laplace and Jeffreys-Carnap
$res
ID si Ni BL (mode) JC (mode) BL (mean) JC (mean)
NA 23 27 0.85      0.88      0.83      0.85
BL (sd) JC (sd) BL (var) JC (var)
0.069   0.067   0.0048   0.0045
$hdi.BL
      rn 0.69% 0.95% 0.99%
lower NA 0.78  0.69  0.63
upper NA 0.91  0.95  0.97
$hdi.JC
      rn 0.69% 0.95% 0.99%
lower NA 0.81  0.72  0.66
upper NA 0.93  0.97  0.98
################################
```

También se puede llamar a las funciones para valores individuales de θ:

```
> # single values
> # theta.quer.BL +/- sigma.BL (studer nimmt 75%)
> # BL
> # calculate without logs
> pbl(theta=0.5, si=si, Ni=Ni, loga=FALSE)
[1] 0.003661215
> # calculate with logs and give out log value
> exp( pbl(theta=0.5, si=si, Ni=Ni, loga=TRUE) )
[1] 0.003661215
> # calculate with logs and give out re-exp non-log value
> pbl(theta=0.5, si=si, Ni=Ni, loga=TRUE, reexp=TRUE)
[1] 0.003661215
>
> # JC
> # calculate without logs
> pjc(theta=0.5, si=si, Ni=Ni, loga=FALSE)
[1] 0.001782179
```

```
> # calculate with logs and give out log value
> exp( pjc(theta=0.5, si=si, Ni=Ni, loga=TRUE) )
[1] 0.001782179
> # calculate with logs and give out re-exp non-log value
> pjc(theta=0.5, si=si, Ni=Ni, loga=TRUE, reexp=TRUE)
[1] 0.001782179
```

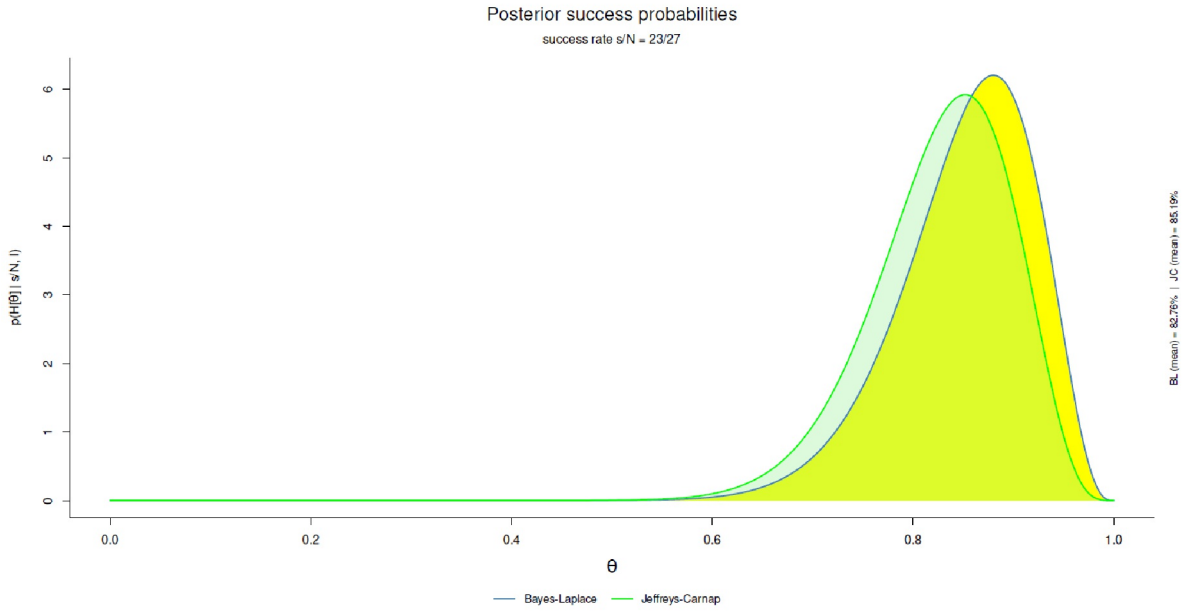


Figura 6.129. *start again*
(porcentajes de éxito, 23 de 27 éxitos, Posterior BL o JC según Studer, 1996b).

Para la cuadrícula, el intervalo válido de valores se denota de 0 a 1. Además, los valores extremos son interesantes, como ya se ha anunciado. El caso extremo para la supuesta sobre-estimación del fracaso es $s_i = 0$ y $N_i = 6$, para el caso de éxito $s_i = 6$ y $N_i = 6$. Cabe señalar que aquí se producen mensajes de error o advertencia, ya que los cálculos de la cuadrícula pueden producir valores que se encuentran fuera del intervalo válido de valores. El motivo son las funciones gamma. Las figuras 6.130, 6.131 y 6.132 muestran los resultados. El aspecto de las probabilidades posteriores se conoce a partir de la distribución beta (véase la Fig. 6.68).

```
# no success / extreme case
si <- 0
Ni <- 6
pbl.res <- pbl(theta=theta, si=si, Ni=Ni, loga=TRUE, reexp=TRUE)
pjc.res <- pjc(theta=theta, si=si, Ni=Ni, loga=TRUE, reexp=TRUE)
plot.bl.jc(theta, sN.ME.res=data.frame(pbl.res, pjc.res), si=si, Ni=Ni, filling=TRUE)
# only success / extreme case
si <- 6
Ni <- 6
pbl.res <- pbl(theta=theta, si=si, Ni=Ni, loga=TRUE, reexp=TRUE)
pjc.res <- pjc(theta=theta, si=si, Ni=Ni, loga=TRUE, reexp=TRUE)
plot.bl.jc(theta, sN.ME.res=data.frame(pbl.res, pjc.res), si=si, Ni=Ni, filling=TRUE)
# one success / less extreme case
si <- 1
Ni <- 6
pbl.res <- pbl(theta=theta, si=si, Ni=Ni, loga=TRUE, reexp=TRUE)
pjc.res <- pjc(theta=theta, si=si, Ni=Ni, loga=TRUE, reexp=TRUE)
plot.bl.jc(theta, sN.ME.res=data.frame(pbl.res, pjc.res), si=si, Ni=Ni, filling=TRUE)
```

```
# almost everything a success / less extreme case
si <- 5
Ni <- 6
pbl.res <- pbl(theta=theta, si=si, Ni=Ni, loga=TRUE, reexp=TRUE)
pjc.res <- pjc(theta=theta, si=si, Ni=Ni, loga=TRUE, reexp=TRUE)
plot.bl.jc(theta, sN.ME.res=data.frame(pbl.res, pjc.res), si=si, Ni=Ni, filling=TRUE)
# not everything a success / less extreme case
si <- 4
Ni <- 6
pbl.res <- pbl(theta=theta, si=si, Ni=Ni, loga=TRUE, reexp=TRUE)
pjc.res <- pjc(theta=theta, si=si, Ni=Ni, loga=TRUE, reexp=TRUE)
plot.bl.jc(theta, sN.ME.res=data.frame(pbl.res, pjc.res), si=si, Ni=Ni, filling=TRUE)
```

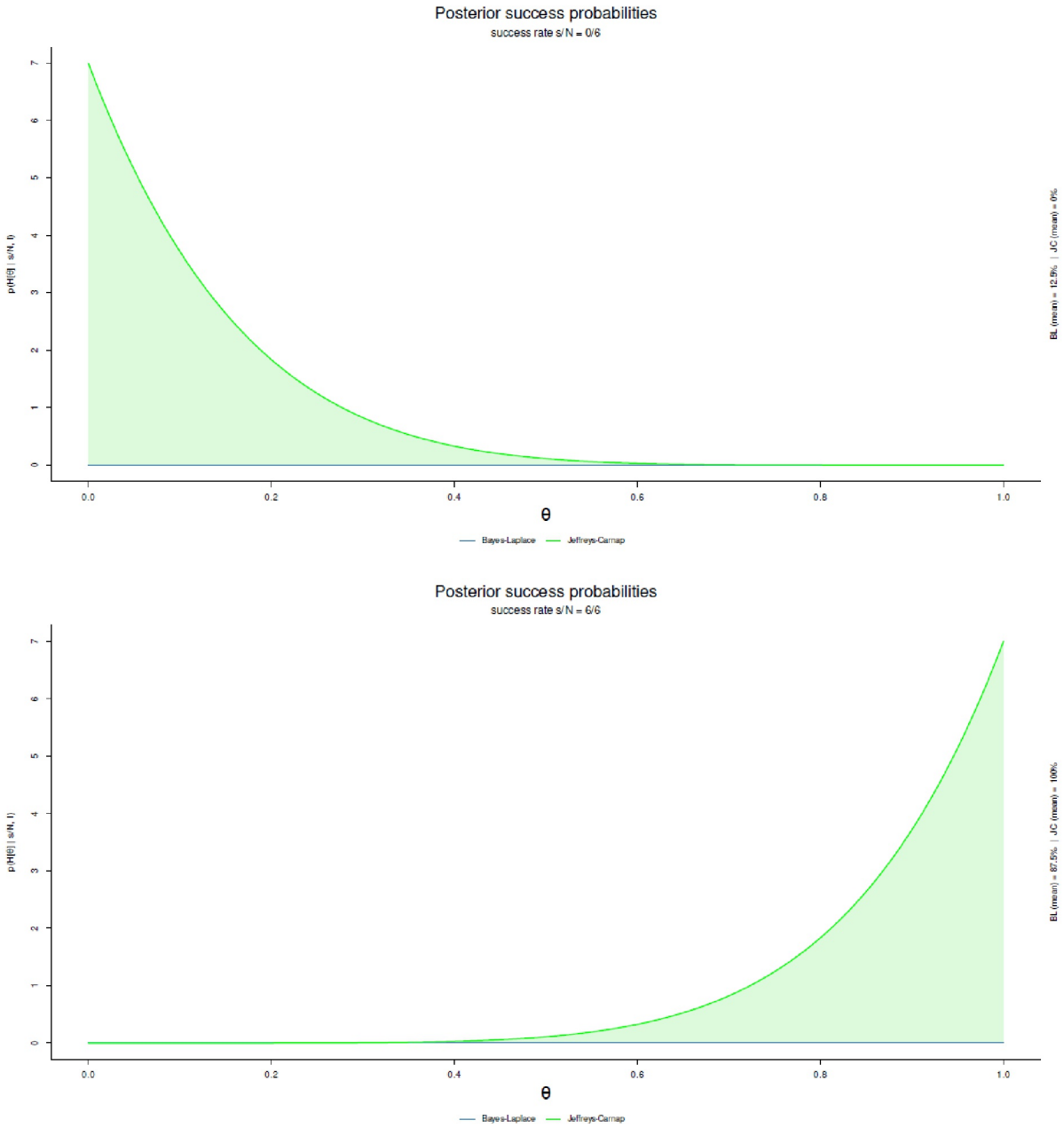


Figura 6.130. *start again*

(porcentajes de aprobados parte 1, arriba: 0 de 5 aciertos, abajo: 6 de 6 aciertos)

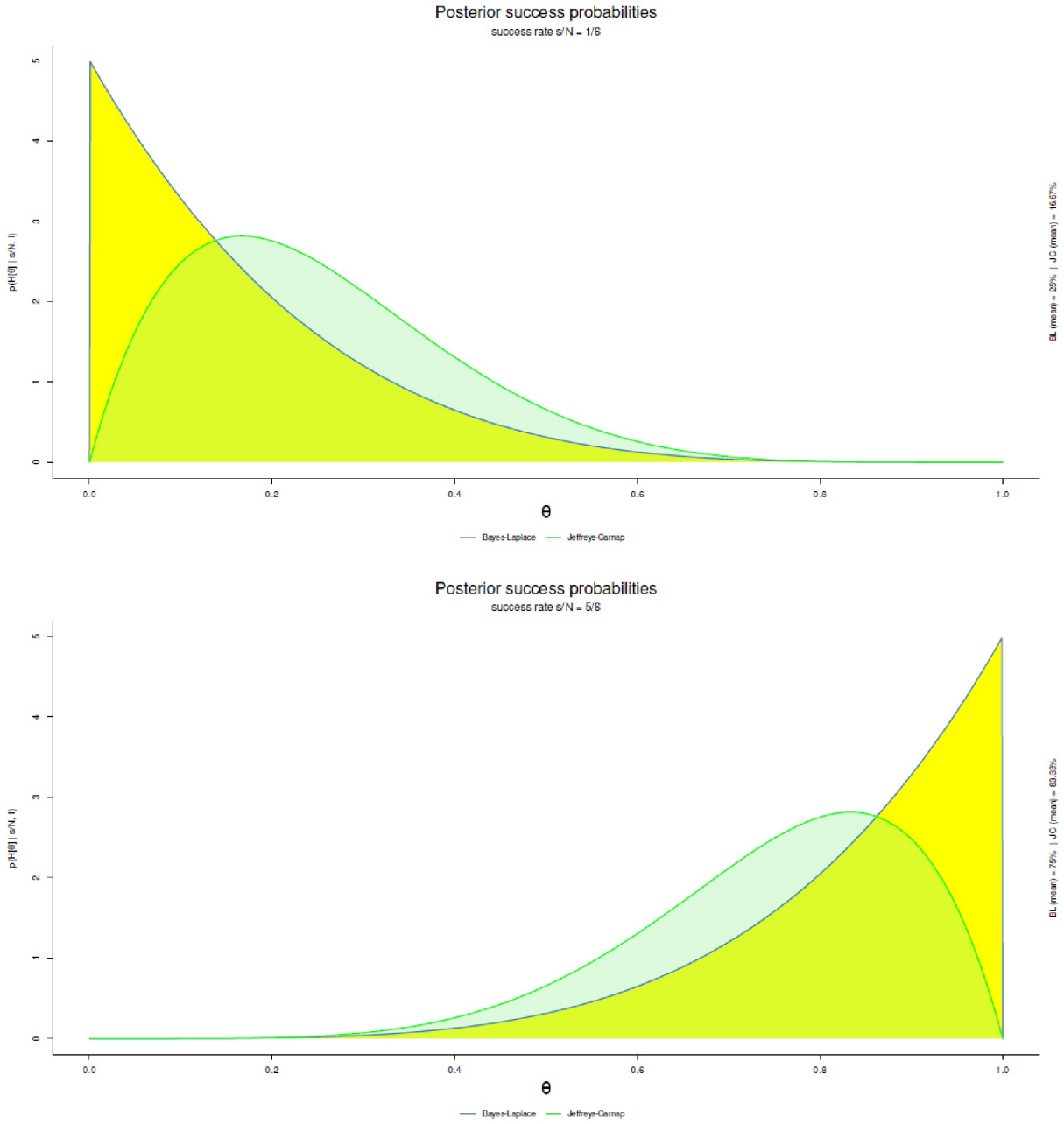


Figura 6.131. *start again*
(porcentajes de aprobados parte 2, arriba: 1 de 6 aciertos, abajo: 5 de 6 aciertos)

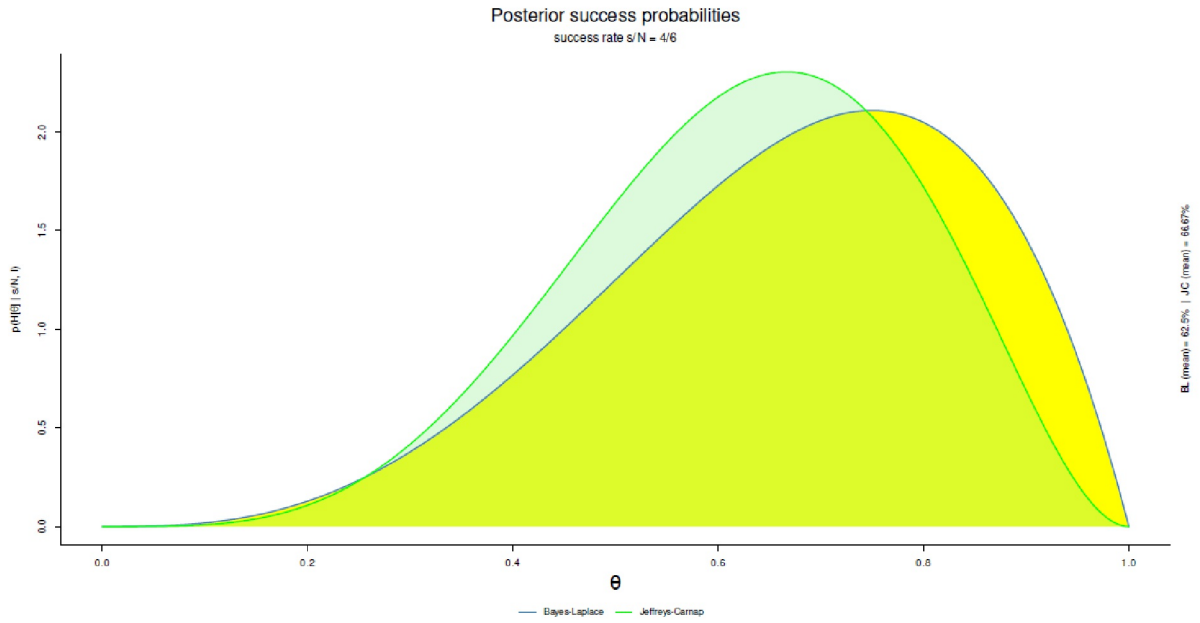


Figura 6.132. *start again (porcentajes de aprobados parte 3, 4 de 6 aciertos)*

6.15.2.1 Evaluación a largo plazo 1992-2017

La situación así preparada puede ampliarse para incluir datos empíricos a largo plazo, a saber, los porcentajes de aprobados de `start again` entre 1992 y 2017. En la tabla 6.13 figuran los datos brutos. Los datos se extrajeron de los informes anuales publicados y se completaron con datos de la dirección, que se pusieron generosamente a nuestra disposición aquí. Los datos a largo plazo permiten reconstruir la evolución, independientemente de las fluctuaciones interanuales (= Likelihood), para trazar la suma de la experiencia a lo largo de los años (= Posterior). La suma de la experiencia corresponde a una estimación robusta, en la que las fluctuaciones anuales influyen cada vez menos a lo largo de los años. Al mismo tiempo, las dos fórmulas anteriores Jeffreys-Carnap con $p_{JC}(H \mid s/N, I)$ y Bayes-Laplace con $p_{BL}(H \mid s/N, I)$ según Studer (1996b), que reflejan un trasfondo previo diferente, sirven directamente de comparación. Esto se puede visualizar a nivel gráfico representando la Prior, la Likelihood y la Posterior en un único gráfico para cada año (véase la Fig. 6.134). Primero se leen los datos (`ptII_quan_Bayes_case_startagain-successrates-longterm.r`).

```
# read source tab R-Code
sa <- read.table("startagain_statistics_1992-2017.tab", sep="\t", header=TRUE)
dim(sa)
sa
```

Dado que los años 1992 (= año de fundación) y 2017 no comprenden 12 meses completos, se tiene que corregir los datos en consecuencia. Para simplificar (con una ligera distorsión) escalamos a valores anuales. También sería posible trabajar exclusivamente con los valores mensuales.

```
# correct for parts of a year (first and last year, ie. 1992 and 2017)
mpy <- 12 # months per year
sa.cor <- cbind(year=sa[,"year"],
       sa[,c("IN","OUT","s","f","indiff")] /
       sa[,"mpyear"]*mpy)
sa.cor
```

Tabla 6.13: *start again (índices de aprobados 1992-2017, datos brutos)*

Año	Movimientos			Outcome			$Outcome_{cum}$			months/ year
	Ni_{cum}	IN	OUT	s(uccess)	f(ailure)	indiff(erent)	$s_{cum.}$	f_{cum}	$indiff_{cum}$	
1992	4	2	2	0	2	0	0	2	0	3
1993	53	49	23	2	21	0	2	23	0	12
1994	92	39	42	15	27	0	17	50	0	12
1995	121	29	32	15	17	0	32	67	0	12
1996	152	31	26	15	11	0	47	78	0	12
1997	182	30	35	19	16	0	66	94	0	12
1998	218	36	38	20	18	0	86	112	0	12
1999	251	33	30	18	12	0	104	124	0	12
2000	278	27	29	11	18	0	115	142	0	12
2001	306	28	31	18	13	0	133	155	0	12
2002	331	25	23	13	10	0	146	165	0	12
2003	352	21	27	17	10	0	163	175	0	12
2004	374	22	17	12	5	0	175	180	0	12
2005	389	15	28	14	8	6	189	188	6	12
2006	412	23	17	11	3	3	200	191	9	12
2007	433	21	20	17	3	0	217	194	9	12
2008	445	12	17	8	7	2	225	201	11	12
2009	472	27	26	9	16	1	234	217	12	12
2010	496	24	20	7	12	1	241	229	13	12
2011	513	17	15	10	5	0	251	234	13	12
2012	526	13	18	13	5	0	264	239	13	12
2013	544	18	18	11	7	0	275	246	13	12
2014	559	15	23	7	14	2	282	260	15	12
2015	569	10	15	6	7	2	288	267	17	12
2016	581	12	12	2	10	0	290	277	17	12
2017	582	1	3	2	1	0	292	278	17	3

En principio, se puede clasificar los seguidores indiferentes – que no completaron el programa por diversos motivos – como abandonos o como éxitos (véase Studer, 1998). La tabla 6.13 recoge los datos brutos de ambas variantes. Para los cálculos de los porcentajes de aprobados (= Posterior) recurrimos a la variante conservadora y los asignamos a los abandonos. No obstante, creamos columnas para el caso conservador y el caso menos conservador. Los lectores interesados podrían repetir el análisis para el caso menos conservador y comparar los porcentajes de éxito con el caso conservador.

```
# conservative: failure = failure + indiff
f.indiff <- rowSums(sa.cor[,c("f","indiff")])
# less conservative: success = success + indiff
s.indiff <- rowSums(sa.cor[,c("s","indiff")])
```

Los datos están disponibles individualmente por año. Por lo tanto, necesitamos los totales por año.

```
# N per year
N <- rowSums(sa.cor[,c("s","f","indiff")])
sa.cor.enh <- data.frame(sa.cor, f.indiff, s.indiff, N)
sa.cor.enh
```

Para los cálculos a partir de la misma Prior, también necesitamos valores acumulados (tasas de aprobados, total de casos). Estos pueden sumarse mediante `cumsum()`.

```
# cumsums
sa.cor.enh.cs <- apply(sa.cor.enh[,-1], 2, cumsum)
colnames(sa.cor.enh.cs) <- paste(colnames(sa.cor.enh.cs),".cs", sep="")
sa.cor.enh.cs
sa.all <- data.frame(sa.cor.enh, sa.cor.enh.cs)
sa.all
```

Los valores para la Posterior se generan por año utilizando `sN.post.su()`. Esta función toma como entrada la tasa de éxito s_i y el número de casos *Ni*. A continuación se generan los HDI y los estadísticos de resumen.

```
# prepare lists
sa.l <- dim(sa.all)
sa.l
res.bino.hdi.EXP <- res.bino.sum.EXP <- res.bino.hdi <-
res.bino.sum <- res.sum <- res.hdi.BL <- res.hdi.JC <- list()
theta.prior <- 0.5
nprior <- 2
# calculate everything
for(i in 1:sa.l[1])
{
#print(i)
temp <- sN.post.su(Ni=sa.all[i,"N.cs"], si=sa.all[i,"s.cs"],
       rn=sa.all[i,"year"], printout=FALSE)
res.sum[[i]] <- temp[["res"]]
res.hdi.BL[[i]] <- temp[["hdi.BL"]]
res.hdi.JC[[i]] <- temp[["hdi.JC"]]
#uniform prior = identical to JC
temp <- bino.abs(si=sa.all[i,"s.cs"], Ni=sa.all[i,"N.cs"],
       theta.prior=theta.prior, nprior=nprior,
       rn=sa.all[i,"year"], graph=FALSE)
res.bino.sum[[i]] <- temp[["res"]]
res.bino.hdi[[i]] <- temp[["hdi"]]
#learning from experience ie. prior = posterior[i-1]
if(i==1)
{
  temp.EXP <- bino.abs(si=sa.all[i,"s.cs"], Ni=sa.all[i,"N.cs"],
  theta.prior=theta.prior, nprior=nprior,
       rn=sa.all[i,"year"], graph=FALSE)
} else
{
  theta.prior <- NULL
  nprior <- NULL
  a.prior <- res.bino.sum.EXP[[i-1]][,"a.post"]
  b.prior <- res.bino.sum.EXP[[i-1]][,"b.post"]
  temp.EXP <- bino.abs(si=sa.all[i,"s"], Ni=sa.all[i,"N"],
       a.prior=a.prior, b.prior=b.prior,
       rn=sa.all[i,"year"], graph=FALSE)
}
res.bino.sum.EXP[[i]] <- temp.EXP[["res"]]
res.bino.hdi.EXP[[i]] <- temp.EXP[["hdi"]]
}
```

Se pueden observar varios mensajes de error. La mayoría de ellos tienen que ver con el hecho de que, o bien en casos extremos (véase más arriba) los valores se encuentran fuera del rango admisible para las funciones gamma o bien para una distribución Prior uniforme (a = 1, b = 1) no es posible ningún HDI. En consecuencia estas celdas se marcan con NA como no presentes. Ignoramos estas celdas porque sabemos cómo se forman.

Tabla 6.14: *start again (tasas de aprobados 1992-2017, JC- o BL-Posterior según Studer, 1996b)*

Año	s	N	Modo		$\bar{x}$		s		s^2	
			BL	JC	BL	JC	BL	JC	BL	JC
1992	0	8	0.0000	NA	0.1000	0.0000	0.0905	0.0000	0.0082	0.0000
1993	2	31	0.0641	0.0340	0.0909	0.0645	0.0493	0.0434	0.0024	0.0019
1994	17	73	0.2332	0.2252	0.2400	0.2329	0.0490	0.0491	0.0024	0.0024
1995	32	105	0.3043	0.3013	0.3084	0.3048	0.0444	0.0447	0.0020	0.0020
1996	47	131	0.3584	0.3564	0.3609	0.3588	0.0415	0.0417	0.0017	0.0017
1997	66	166	0.3974	0.3964	0.3988	0.3976	0.0377	0.0379	0.0014	0.0014
1998	86	204	0.4214	0.4204	0.4223	0.4216	0.0343	0.0345	0.0012	0.0012
1999	104	234	0.4444	0.4444	0.4449	0.4444	0.0323	0.0324	0.0010	0.0011
2000	115	263	0.4374	0.4364	0.4377	0.4373	0.0304	0.0305	0.0009	0.0009
2001	133	294	0.4525	0.4525	0.4527	0.4524	0.0289	0.0290	0.0008	0.0008
2002	146	317	0.4605	0.4605	0.4608	0.4606	0.0279	0.0280	0.0008	0.0008
2003	163	344	0.4735	0.4735	0.4740	0.4738	0.0268	0.0269	0.0007	0.0007
2004	175	361	0.4845	0.4845	0.4848	0.4848	0.0262	0.0263	0.0007	0.0007
2005	189	389	0.4855	0.4855	0.4859	0.4859	0.0252	0.0253	0.0006	0.0006
2006	200	406	0.4925	0.4925	0.4926	0.4926	0.0247	0.0248	0.0006	0.0006
2007	217	426	0.5095	0.5095	0.5093	0.5094	0.0241	0.0242	0.0006	0.0006
2008	225	443	0.5075	0.5075	0.5079	0.5079	0.0237	0.0237	0.0006	0.0006
2009	234	469	0.4985	0.4985	0.4989	0.4989	0.0230	0.0231	0.0005	0.0005
2010	241	489	0.4925	0.4925	0.4929	0.4928	0.0225	0.0226	0.0005	0.0005
2011	251	504	0.4985	0.4985	0.4980	0.4980	0.0222	0.0222	0.0005	0.0005
2012	264	522	0.5055	0.5055	0.5057	0.5057	0.0218	0.0219	0.0005	0.0005
2013	275	540	0.5095	0.5095	0.5092	0.5093	0.0215	0.0215	0.0005	0.0005
2014	282	563	0.5005	0.5005	0.5009	0.5009	0.0210	0.0211	0.0004	0.0004
2015	288	578	0.4985	0.4985	0.4983	0.4983	0.0207	0.0208	0.0004	0.0004
2016	290	590	0.4915	0.4915	0.4916	0.4915	0.0205	0.0206	0.0004	0.0004
2017	298	602	0.4955	0.4955	0.4950	0.4950	0.0203	0.0204	0.0004	0.0004

```
Werte = NaN bei theta = 0
Several or no maximum (MAP/ mode) for Jeffreys-Carnap.
Plot the function with these parameters to understand 'why'.
s = 0 of N = 8 -> i.e. no successes ->
HPD interval does not make sense for Bayes-Laplace and
Jeffreys-Carnap method (success rates).
Plot the function with these parameters to understand 'why'.
s = 0 of N = 8 -> i.e. not enough successes or not enough
data (= N) -> HPD interval does not make sense for
Jeffreys-Carnap method (success rates).
Plot the function with these parameters to understand 'why'.
```

Se puede reensamblar las listas individuales en tablas con `do.call("rbind", ...)`.

```
# create tables
sa.res <- do.call("rbind",res.sum)
sa.hdi.BL <- do.call("rbind",res.hdi.BL)
sa.hdi.JC <- do.call("rbind",res.hdi.JC)
sa.bino <- do.call("rbind",res.bino.sum)
sa.bino.EXP <- do.call("rbind",res.bino.sum.EXP)
sa.bino.hdi.EXP <- do.call("rbind",res.bino.hdi.EXP)
```

En la tabla 6.14 se enumeran los valores anuales de las Posteriores multiplicando cada año sobre la base de la Likelihood acumulada (es decir, todos los valores hasta el año inclusive) por las distribuciones a priori *JC* y *BL* según Studer (1996b, véase más arriba). Esto *no* corresponde al caso sucesivo de *aprendizaje a partir de la experiencia*, sino que supone – ficticiamente – por cálculo anual de el mismo nivel de conocimientos previos *JC* o *BL*, pero sobre la base de datos acumulados. Dado que el proceso generador de datos es un modelo binomial (= Likelihood), es decir, éxito y fracaso, la distribución beta (puesto que se trata de Priores conjugadas) se presta a describir la Prior y la Posterior por igual. A continuación, la Figura 6.133 visualiza los cambios en la media, la desviación estándar y la proporción de estos coeficientes en función de *JC* y *BL*, respectivamente. Como puede observarse, las curvas para *JC* y *BL* se solapan casi por completo tras diferentes evoluciones a corto plazo.

```
# plot change/ development of mean, sd, and ratio mean/sd
sN.sum.plot(tab=sa.res,
       TITLE="Development success rates (start again 1992-2017)",
       xlab="year", type="l")
```

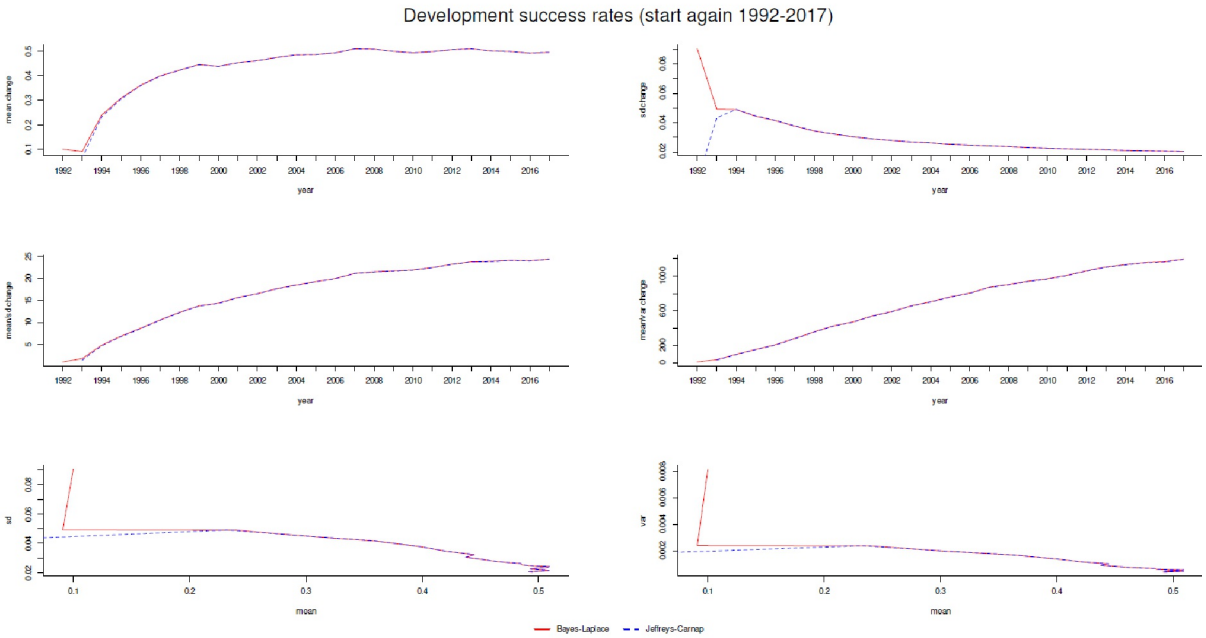


Figura 6.133. *start again* (*tasas de aprobados 1992-2017,* μ, σ, μ/σ)

La tabla 6.15 añade el proceso de *aprendizaje a partir de la experiencia*. En este caso, los parámetros posteriores de la distribución beta del año en curso se toman como parámetros de entrada para la Prior del año siguiente, de modo que se traza la acumulación gradual de experiencia (véase también la Fig. 6.134). *Los conocimientos de ayer constituyen el punto de partida de hoy*. Mientras que al principio la probabilidad determina fuertemente la Posterior, esta influencia de los datos anuales disminuye y la Prior – en este caso la riqueza de la experiencia de los años anteriores, es decir, la Posterior del año anterior – determina cada vez más la aparición de la Posterior actual. Aprender de la experiencia para estos datos significa aprender de los datos anteriores, no aprender a partir de opiniones subjetivas. La probabilidad representa los nuevos

datos anuales, mientras que la anterior incluye todos los datos anteriores. Sólo en el primer año, cuando aún no se dispone de datos empíricos, es necesaria una Prior no empírica y tiene un impacto. Esta influencia se desvanece en favor de los datos empíricos en los años siguientes. Es importante entender correctamente estos dos enfoques diferentes. Si se pretende no saber nada cada año y no se aprende del pasado – es decir, tomar la misma Prior uniforme cada año pero trabajar con datos acumulados (véase más arriba) – la probabilidad define en gran medida la Posterior y sólo en pequeña medida la Prior. Sin embargo, si se lleva a cabo un proceso de actualización anual, la Prior, como la Posterior del año anterior, asume la tarea de determinar la Posterior y la Likelihood sólo corrige a lo largo de la evolución del año anterior. Este proceso se determina de forma puramente empírica, ya que la Prior representa el conocimiento empírico y no el conocimiento experto subjetivamente discutible.

Tabla 6.15: *start again (tasas de aprobados 1992-2017, actualización beta con Prior, Likelihood y Posterior, aprendizaje de la experiencia)*

Año	Prior		Likelihood		Posterior					
	a	b	a	b	a	b	Modus	$\bar{x}$	s	s^2
1992	1	1	1	9	1	9	0.0000	0.1000	0.0905	0.0082
1993	1	9	3	22	3	30	0.0645	0.0909	0.0493	0.0024
1994	3	30	16	28	18	57	0.2329	0.2400	0.0490	0.0024
1995	18	57	16	18	33	74	0.3048	0.3084	0.0444	0.0020
1996	33	74	16	12	48	85	0.3588	0.3609	0.0415	0.0017
1997	48	85	20	17	67	101	0.3976	0.3988	0.0377	0.0014
1998	67	101	21	19	87	119	0.4216	0.4223	0.0343	0.0012
1999	87	119	19	13	105	131	0.4444	0.4449	0.0323	0.0010
2000	105	131	12	19	116	149	0.4373	0.4377	0.0304	0.0009
2001	116	149	19	14	134	162	0.4524	0.4527	0.0289	0.0008
2002	134	162	14	11	147	172	0.4606	0.4608	0.0279	0.0008
2003	147	172	18	11	164	182	0.4738	0.4740	0.0268	0.0007
2004	164	182	13	6	176	187	0.4848	0.4848	0.0262	0.0007
2005	176	187	15	15	190	201	0.4859	0.4859	0.0252	0.0006
2006	190	201	12	7	201	207	0.4926	0.4926	0.0247	0.0006
2007	201	207	18	4	218	210	0.5094	0.5093	0.0241	0.0006
2008	218	210	9	10	226	219	0.5079	0.5079	0.0237	0.0006
2009	226	219	10	18	235	236	0.4989	0.4989	0.0230	0.0005
2010	235	236	8	14	242	249	0.4928	0.4929	0.0225	0.0005
2011	242	249	11	6	252	254	0.4980	0.4980	0.0222	0.0005
2012	252	254	14	6	265	259	0.5057	0.5057	0.0218	0.0005
2013	265	259	12	8	276	266	0.5093	0.5092	0.0215	0.0005
2014	276	266	8	17	283	282	0.5009	0.5009	0.0210	0.0004
2015	283	282	7	10	289	291	0.4983	0.4983	0.0207	0.0004
2016	289	291	3	11	291	301	0.4915	0.4916	0.0205	0.0004
2017	291	301	9	5	299	305	0.4950	0.4950	0.0203	0.0004

Encontramos la salida así:

```
# demonstrate changes in prior, likelihood, posterior R-Code
# start again data 1992-2017
sa.all
sa.res
```

Veamos ahora el código R. La actualización de los valores de la distribución beta sigue las explicaciones del capítulo 6.12.1. Comienza con la conversión de las tasas de éxito *s* y el número total de casos *N* en los parámetros *a* y *b* de la distribución beta.

```
# pre year
sa.res.py <- matrix(data=NA, nrow=dim(sa.res)[1], ncol=6,
dimnames=list(rownames(sa.res),
c("a.prior","b.prior","a.lik","b.lik","a.post","b.post"))
)
sa.res.py <- data.frame(sa.res.py)
head(sa.res.py)
dim(sa.res.py)
# 1,1 = uniform
# .5,.5 = Jeffrey's
# .35 = 7,20 = informed weaker success prior
# .35*20=7
# .55 = 11,20 = informed stronger success prior
# .55*20 = 11
# .75 = 15,20 = unreal stronger success prior
# .75*20 = 15
# total domination: 1900,2000
# sa.res.py[1,"a.prior"] <- 1
# sa.res.py[1,"b.prior"] <- 1
ab.likelis <- do.call("cbind", bino.ab.lik(sa.all["s"],sa.all["N"]))
ab.likelis
sa.res.py[,c("a.lik","b.lik")] <- ab.likelis
sa.res.py
```

A continuación se elige la Prior. De nuevo, se pasan los parámetros *a* y *b* de la distribución beta.

```
# choose different priors
sa.res.py[1,c("a.prior","b.prior")] <- c(1,1)
sa.res.py
sa.res.py[1,c("a.prior","b.prior")] <- c(0.01,0.01)
sa.res.py
sa.res.py[1,c("a.prior","b.prior")] <- c(0.5,0.5)
sa.res.py
```

Los cálculos alternativos de Posterior y Prior consisten únicamente en la aplicación de las fórmulas comunes (Bolstadt, 2007, cap. 8). Es posible elegir si se desea la misma Prior (véase más arriba) o si se desea la alternancia de Prior y Posterior. Como resultado es una tabla con los parámetros *a* y *b* de la Prior, Likelihood y Posterior. Esto también es posible para la Likelihood, una distribución binomial (proceso de Bernoulli), ya que las tasas de éxito y el número total de casos se pueden expresar como los parámetros *a* y *b* de una distribución beta (véase el código R).

```
# calculate following priors and posteriors
# same prior
sa.res.sp <- betabinomial.lbyxp(sa.res.py=sa.res.py,
        prior=list(a=1,b=1),
        sameprior=TRUE)
# prior[i] = posterior[i-1]
sa.res.dp <- betabinomial.lbyxp(sa.res.py=sa.res.py,
        prior=list(a=1,b=1))
# collect and merge infos
abc <- do.call("cbind", beta.summary(a=sa.res.dp[,"a.post"],
```

```
        b=sa.res.dp[,"b.post"]))
colnames(abc) <- c("a.post","b.post","mode.post",
        "mean.post","sd.post","var.post")
res.sa <- data.frame(sa.all, sa.res.dp, abc)
```

Miramos la salida:

```
# same prior R-Code
sa.res.sp
# different prior
sa.res.dp
# all results
res.sa
```

Se puede ampliar esto para incluir estadísticas de resumen y representarlas gráficamente, ya sea en un gran gráfico o de forma consecutiva con la función de R `beta.triplot()`.

```
# one plot after each other R-Code
par(ask=TRUE)
for(i in 1:26)
{
beta.triplot(si=res.sa[i,"s"], Ni=res.sa[i,"N"], v=res.sa[i,],
        multiplot=FALSE, rn=res.sa[i,"year"])
}
# everything on one plot
par(mfrow=c(5,6)) # not 6,5 -> margins too large... error
for(i in 1:26)
{
beta.triplot(si=res.sa[i,"s"], Ni=res.sa[i,"N"], v=res.sa[i,],
        multiplot=TRUE, rn=res.sa[i,"year"])
}
plot.new()
legend("center",legend=c("prior","likelihood","posterior"),
        xpd=TRUE, horiz=FALSE, inset=c(0,0), y.intersp=1,
        col=c("blue","green","red"), lty=c(2,3,1), lwd=1.9, bty="n", cex=2)
plot.new()
mtext(side=1, "Development\nsuccess rates\nstart again\n\n1992-2017", cex=1.2)
```

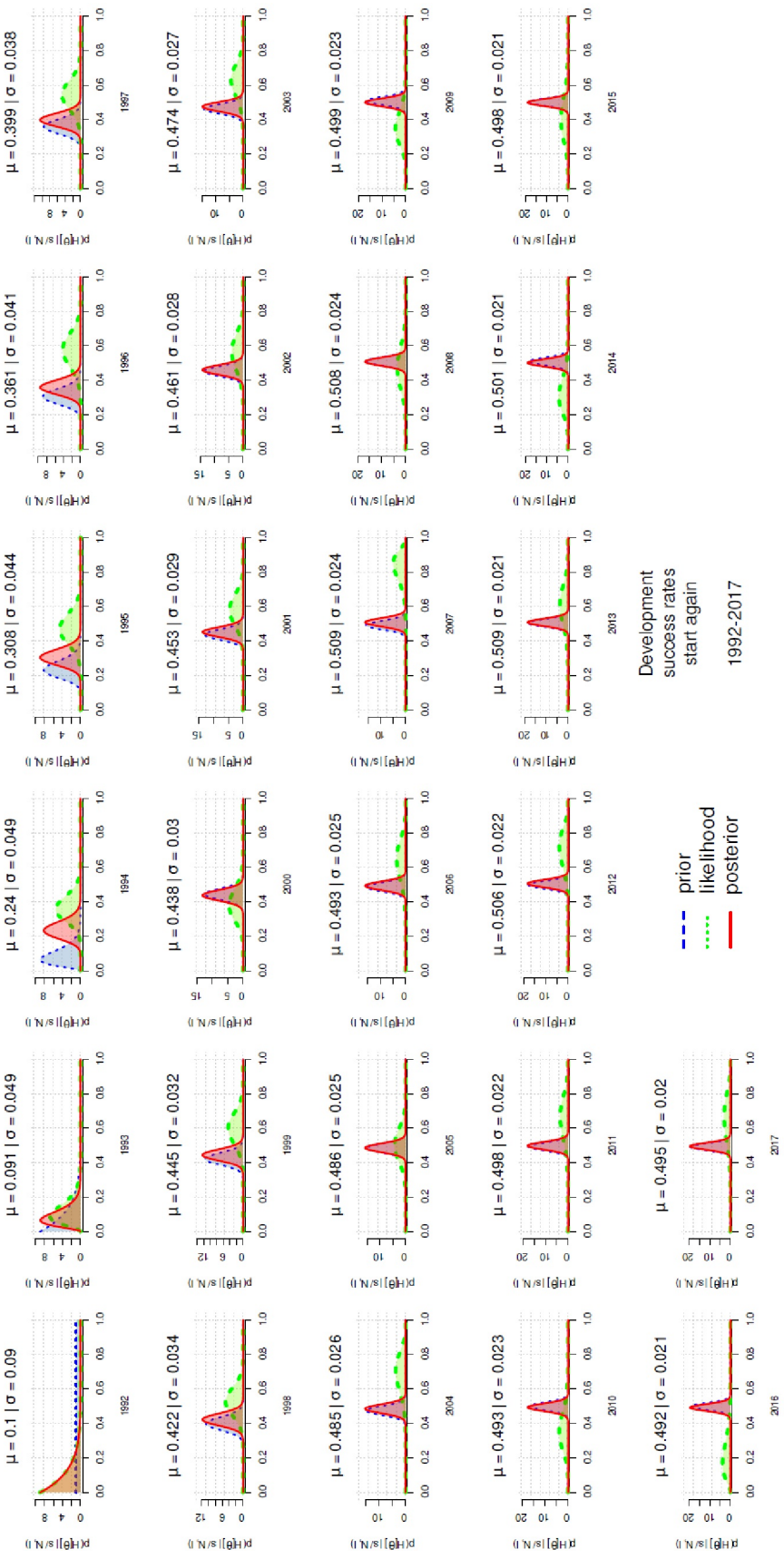


Figura 6.134. *start again* (*tazas de aprobados 1992-2017, aprender de la experiencia)*

Si observamos los gráficos (véase la Fig. 6.134) uno tras otro, enseguida queda claro lo que significan en la práctica las observaciones anteriores sobre el aprendizaje de la experiencia. Al principio, la Prior y la Likelihood ejercen una gran influencia sobre la Posterior. La Posterior siempre se mueve entre las dos. Con el paso de los años la Prior y la Posterior se acercan cada vez más, ya que la Prior contiene el conocimiento experiencial acumulado. La Likelihood añade aquí fluctuaciones anuales, pero éstas no afectan demasiado ni a la desviación estándar ni a la media de la Posterior.

6.15.3 Estudio de caso: "Yo, nosotros y la nación" – autopromoción presidencial Parte 2

Seguimos aquí con el análisis del conjunto de datos sobre el debate presidencial entre Bush y Kerry, que se examinó en el capítulo 6.14.3 desde el punto de vista de la entropía. Ahora nos interesa saber si las respectivas distribuciones de los términos son comparables o si existen diferencias estadísticas. La comparación de estas diferencias (véase la Tabla 6.12) permite un aspecto interesante de la estadística de Bayes, a saber, la comparación de la simulación MCMC, la fuerza bruta, la aproximación por cuadrícula y la solución analíticamente exacta, así como los factores de Bayes. Esto permite reproducir el tema de los procedimientos equivalentes – en este caso de análisis de datos – en un conjunto de datos concreto y discutirlo como en el capítulo 4.4.9. Todos estos procedimientos están disponibles para el análisis de tablas de frecuencias o proporciones binomialmente distribuidas. Se describen los procedimientos en forma de esquema. De modo bayesiano examinamos las siguientes variantes (`ptII_quan_Bayes_caso_debates_-presidenciales.r`):

- MCMC estimación vía JAGS (s. cap. 6.15.3.1)
- MCMC vía fuerza bruta (en dos variantes, s. cap. 6.15.3.2 y 6.15.3.3)
- Integración numérica o aproximación mediante un cuadrículo (véase el capítulo 6.15.3.4)
- Solución analítica exacta (según Evan Miller, 2015, véase el cap. 6.15.3.5)
- Solución analítica exacta (según Pham-Gia, Turkkan y Eng (1993 o Nadarajah y Kotz, 2007, véase la sección 6.136)
- Factor de Bayes (s. cap. 6.15.3.7)
- Prueba *t* bayesiana (según Bretthorst, 1993, véase el capítulo 6.15.3.8)

6.15.3.1 Estimación MCMC mediante JAGS

Los datos (véase la Tabla 6.12) están disponibles en forma resumida en los tres duelos orales para los términos "yo", "nosotros" y "nación".

```
> # start the presidential debates analyses
> pres <- t(matrix(c(16,101,91, 32,131,88), nrow=2, byrow=TRUE,
dimnames=list(c("Bush","Kerry"),c("nation","I","we"))))
> pres
       Bush Kerry
nation  16   32
I      101  131
we      91   88
> counts.bk <- c(16,101,91, 32,131,88)
> president <- gl(2,3, labels=c("Bush","Kerry"))
> term <- gl(3,1,6, labels=c("nation","I","we"))
> pres.dat <- data.frame(counts.bk,president,term)
```

```
> pres.dat
counts.bk president term
1 16  Bush nation
2 101 Bush I
3 91  Bush we
4 32  Kerry nation
5 131 Kerry I
6 88  Kerry we
```

Añadimos las sumas marginales.

```
> # descriptive
> # rows> prop.table(pres, 1)
       Bush      Kerry
nation 0.3333333 0.6666667
I      0.4353448 0.5646552
we     0.5083799 0.4916201
> margin.table(pres, 1)
nation I   we
48     232 179
> # cols
> prop.table(pres, 2)
       Bush       Kerry
nation 0.07692308 0.1274900
I      0.48557692 0.5219124
we     0.43750000 0.3505976
> margin.table(pres, 2)
Bush Kerry
208  251
> # all
> addmargins(pres)
       Bush Kerry Sum
nation  16   32    48
I      101  131   232
we      91   88   179
Sum    208  251   459
```

El punto de partida es primero la solución frecuentista, aquí como una prueba de χ^2 *campos múltiples* con `chisq.test()` o prueba de proporción `prop.test()`.

```
> # frequentist solution
>
> cor(pres[,1], pres[,2])
[1] 0.9427224
> chisq.test(pres)
Pearson's Chi-squared test
data: pres
X-squared = 5.2809, df = 2, p-value = 0.07133
> set.seed(88772)
> chisq.test(pres, sim=TRUE, B=1e5)
Pearson's Chi-squared test with simulated p-value
(based on 1e+05 replicates)
data: pres
X-squared = 5.2809, df = NA, p-value = 0.07311
> prop.test(pres)
3-sample test for equality of proportions
without continuity correction
data: pres
X-squared = 5.2809, df = 2, p-value = 0.07133
alternative hypothesis: two.sided
```

```
sample estimates:
prop 1    prop 2    prop 3
0.3333333 0.4353448 0.5083799
```

Todas soluciones muestran, con $p = 0.071$ desde un punto de vista estrictamente convencional, ningún rechazo estricto de la hipótesis nula a un nivel ligeramente cambiado $\alpha = 7\%$ en lo que respecta al diferente uso de los términos por parte de Bush y Kerry. La tabla puede reducirse aún más contrayendo "nosotros" y "nación" juntos. De este modo podemos reducir a lo esencial en el sentido binario de éxito frente a fracaso (modelo binomial) y repetir lo anterior.

```
# we reduce to 2x2 Chi^2 table R-Code
pres
pres.2x2 <- rbind(pres["I",],pres["nation",]+pres["we",])
rownames(pres.2x2) <- c("I","we/nation")
pres.2x2
addmargins(pres.2x2)
cor(pres.2x2[,1], pres.2x2[,2])
chisq.test(pres.2x2)
set.seed(88772)
chisq.test(pres.2x2, sim=TRUE, B=1e5)
prop.test(pres.2x2)
```

El valor *p* como propiedad de los datos cambia en la dirección de $p = 0.496$. Ahora nos fijamos en la potencia. Con `pwr.2p2n.test()` del paquete `pwr` de R, se puede calcular la potencia para dos proporciones con diferentes tamaños de muestra. Las llamadas siguen las funciones de potencia (véase el capítulo 4.3.3.1).

```
> # power
> # power = ?
> pwr.2p2n.test(h=0.2,n1=208,n2=251,power=NULL,sig=0.07)$power
[1] 0.625971
> # effect size = ?
> pwr.2p2n.test(h=NULL,n1=208,n2=251,power=0.7,sig=0.07)$h
[1] 0.2190626
> # sample size = ?
> pwr.2p2n.test(h=0.2,n1=208,n2=NULL,power=0.7,sig=0.07)$n2
[1] 396.6944
> pwr.2p2n.test(h=0.2,n1=NULL,n2=251,power=0.7,sig=0.07)$n1
[1] 299.0019
```

Suponiendo un tamaño del efecto pequeño habitual de $d = 0.2$ y un nivel de $\alpha = 0.07$, la potencia es de 0.63; podría ser peor. Suponiendo una potencia de 0.7 o β-índice de error de $\beta = 1 - Power = 0.3$, hay un tamaño del efecto pequeño de $d = 0.22$. La cuestión del tamaño de muestra necesario es: si tomamos los datos de Bush como referencia, necesitamos $N_2 = 397$ observaciones para Kerry y viceversa para Bush, dado que Kerry se da, $_{N1 = 299}$ observaciones para poder trabajar (en el futuro) en el marco de los parámetros mencionados.

```
> # necessary samples versus empirical samples (Bush, Kerry)
> c(299,397)/colSums(pres.2x2)
Bush     Kerry
1.437500 1.581673
```

Para Bush necesitaríamos 1.44 veces más observaciones que las dadas y para Kerry 1.58 veces más. Así pues, los datos tienden a ser poco potentes dentro de los parámetros elegidos. Aquí no mostramos curvas. Si está interesado, puede crearlas fácilmente usted mismo ejecutando `pwr.2p2n.test()` sobre diferentes

parámetros (por ejemplo, tamaño de la muestra) y calculando la cantidad de interés (por ejemplo, potencia, tamaño del efecto) en cada caso y luego trazando el resultado. La prueba de proporción bayesiana para tablas de cuatro campos `bayes.prop.test()` utilizando JAGS del paquete `BayesianFirstAid` proporciona el siguiente resultado. La prueba asume una Prior poca informada con Beta(1, 1). Si desea cambiar esto, tiene que modificar el código R y el modelo JAGS (véase más adelante). El valor de comparación para probar la diferencia entre las dos clases de términos por defecto es 0. Esto también se puede cambiar (ver más abajo).

```
# analyze for rows means to t(pres.2x2)
# so that Bush versus Kerry is analyzed!!!
t(pres.2x2)
# rows = terms -> analyze for different frequency of usage
# of the combined terms I vs. we/nation
pres.2x2
# rows = terms -> analyze for different frequency of usage
# between Bush and Kerry
t(pres.2x2)
pres.2x2.bprop <- bayes.prop.test(t(pres.2x2), cred.mass=0.95,
        n.iter=15000, progress.bar="text")
pres.2x2.bprop
# summary
BFA.summary.bayes_prop_test(pres.2x2.bprop)
```

`BFA.summary.bayes_prop_test()` maneja la salida y ...

```
> BFA.summary.bayes_prop_test(pres.2x2.bprop)
Data
Successes Failures N (total)
Group 1 | 101 107 208
Group 2 | 131 120 251
Model parameters and generated quantities
*- theta[i] = the relative frequency of column
'Successes' for Group i
*- x_pred[i] = predicted number of successes
in a replication for Group i
*- theta_diff[i,j] = the difference between two groups
(theta[i] - theta[j])
Measures
                 mean   sd      HDIlo   HDIup %<comp %>comp
theta[1]         0.486  0.035   0.419   0.554 0.657  0.343
theta[2]         0.523  0.031   0.459   0.582 0.241  0.759
x_pred[1]      101.032 10.254  79.000 119.000 0.000  1.000
x_pred[2]      131.117 11.168 109.000 152.000 0.000  1.000
theta_diff[1,2] -0.037  0.047  -0.126   0.055 0.785  0.215
*- 'HDIlo', 'HDIup' = limits of a 95% HDI credible interval.
*- '%<comp', '%>comp' = probabilities of the respective parameter
being smaller or larger than 0.5 (except
for the theta_diff parameters where the
comparison value 'comp' = 0.0).
Quantiles
                q2.5%   q25%    median  q75%    q97.5%
theta[1]          0.419   0.462   0.486   0.509   0.554
theta[2]          0.460   0.501   0.523   0.544   0.584
x_pred[1]        81.000  94.000 101.000 108.000 121.000
x_pred[2]       109.000 123.000 131.000 139.000 153.000
theta_diff[1,2]  -0.127  -0.069  -0.037  -0.005   0.054
Probabilities and Odds Ratios in relation to ROPE [crit < 0.05]
Probability, that the groups (sets) are
equivalent/ the same = 0.58
Probability, that the groups (sets) are
different/ not the same = 0.42
Odds ratio in favor of the groups (sets)
```

```
being equivalent/ the same = 1.382
Odds ratio in favor of the groups (sets)
being different/ not the same = 0.724
Probabilities and Odds Ratios (theta_1 > theta_2)
Probability, that theta_1 > theta_2 = 0.214
Probability, that theta_1 < theta_2 = 0.786
Odds ratio in favor of theta_1 > theta_2 = 0.273
Odds ratio in favor of theta_1 < theta_2 = 3.663
```

... contiene información sobre los datos de referencia, las estimaciones, los cuantiles y las probabilidades posteriores, así como los odds ratios posteriores. Una desviación básica de la diferencia de grupo con respecto a 0 tiene una probabilidad posterior de 0.217. La probabilidad posterior según la hipótesis ROPE de $crit_{equivalencia} < 0.05$ de que los grupos sean iguales es mayor, con $p_{same} = 0.574$, que la de que son diferentes – $p_{di} = 0.426$. La diferencia entre los grupos con $\theta_{diff} =$ 0.036 es más bien pequeña comparada con las medias $\theta_1 = 0.486$ y $\theta_2 = 0.522$, respectivamente, y desviaciones estándar de $\sigma_{\theta 1} = 0.035$ y $\sigma_{\theta 2} = 0.035$, respectivamente.

La figura 6.135 muestra un diagrama de dispersión de los parámetros estimados θ_1 y θ_2 de los dos grupos. La diagonal discontinua marca la línea 0 perfecta. Las líneas elípticas circundantes representan los distintos HDI.

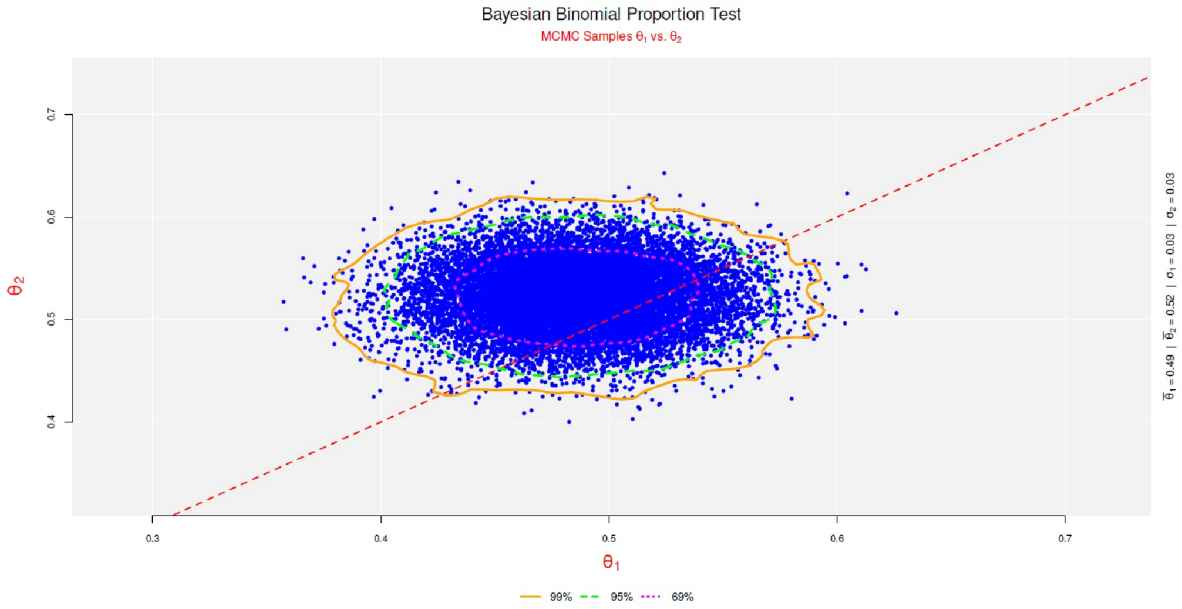


Figura 6.135. *I, we y nation* (*prueba de proporción de Bayes, diagrama de dispersión* θ_1 *y* θ_2)

```
# graphical diagnostics
BFA.mcmcplot.thetas(pres.2x2.bprop)
```

Se puede complementar esto con un gráfico de dispersión con mapa térmico integrado mediante `qplot()` del paquete R `ggplot2` (véase la Fig. 6.136). Esto no hace nada diferente, sólo utiliza un tipo diferente de representación gráfica.

```
qplot(as.data.frame(pres.2x2.bprop)[,1], as.data.frame(pres.2x2.bprop)[,2],
      geom=c("hex"), xlab=expression(theta[1]), ylab=expression(theta[2]))
```

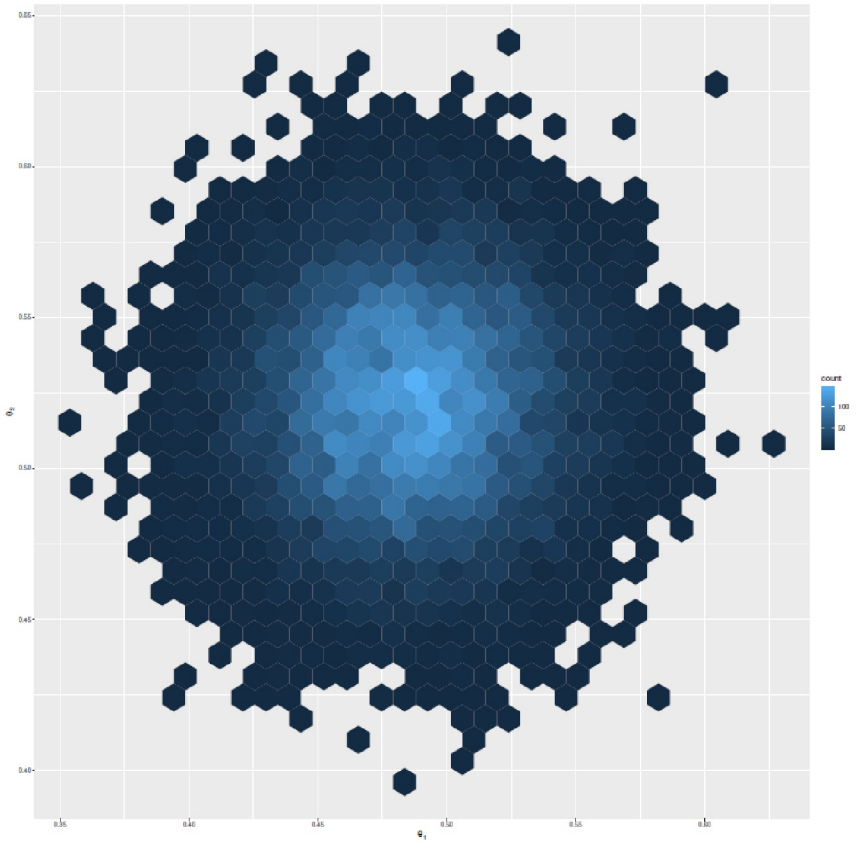


Figura 6.136. I, we y nation
(prueba de proporción de Bayes, *Diagrama de dispersión* θ_1 *y* θ_2 *con mapa térmico)*

Lo que se puede hacer es repetir el análisis. Esto muestra que las probabilidades posteriores de las comparaciones de grupos y la odds ratios cambian ligeramente, lo que se debe a la estimación MCMC de los parámetros. Lo hemos observado a menudo en el contexto de los análisis bayesianos. No se trata de un error, sino que forma parte de la estimación probabilística mediante cadenas de Markov y simulación de Monte Carlo, en este caso con el muestreador de Gibbs (véase el capítulo 6.13.1.2). Comparando las estimaciones de los parámetros, se puede observar que son muy estables a lo largo de diferentes ejecuciones. Reconstruir esto sería una tarea para los lectores dedicados. El examen de las cadenas MCMC no indica ninguna característica especial (véase la Fig. 6.137).

```
# MCMC diagnostics
par(oma=c(2,1,4,1), "cex.axis"=1, bty="l")
diagnostics(pres.2x2.bprop)
mtext("MCMC diagnostics", outer=TRUE, line=1.7, cex=1.5, side=3)
```

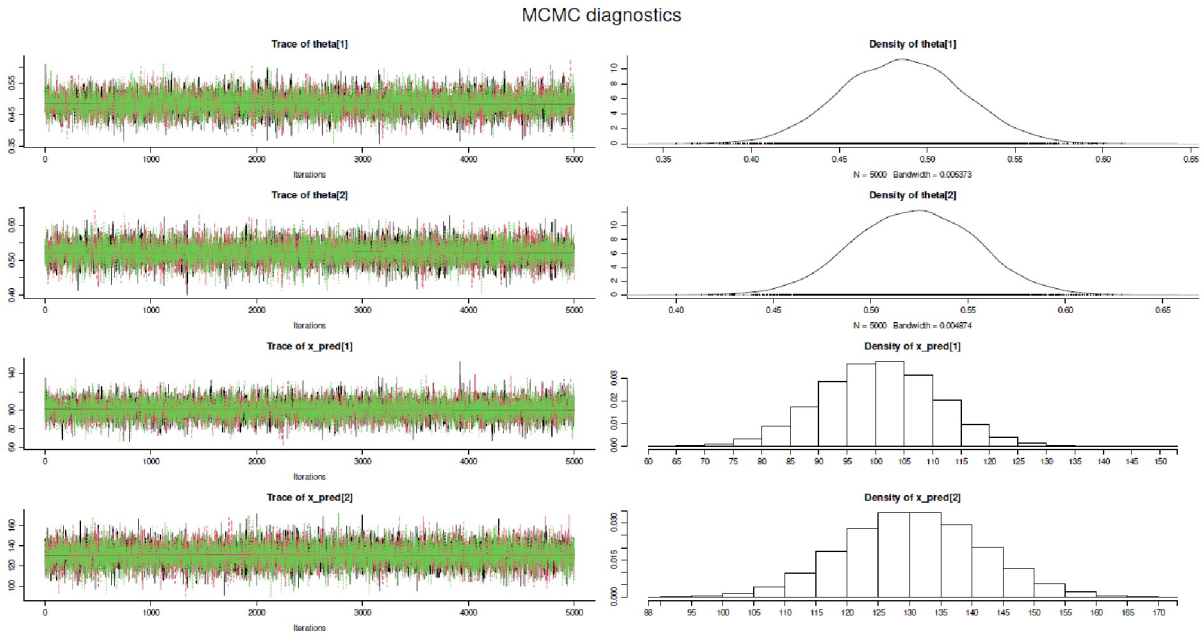


Figura 6.137. *I, we y nation (prueba de proporción de Bayes mediante JAGS, diagnóstico MCMC)*

Ahora un vistazo al código JAGS y la llamada de R:

```
> # model code
> model.code(pres.2x2.bprop)
### Model code for the Bayesian First Aid ###
### alternative to the test of proportions ###
require(rjags)
# Setting up the data
x <- c(Bush = 101, Kerry = 131)
n <- c(Bush = 208, Kerry = 251)
# The model string written in the JAGS language
model_string <- "model {
for(i in 1:length(x)) {
x[i] ~ dbinom(theta[i], n[i])
theta[i] ~ dbeta(1, 1)
x_pred[i] ~ dbinom(theta[i], n[i])
}
}"
# Running the model
model <- jags.model(textConnection(model_string),
data = list(x = x, n = n),
n.chains = 3, n.adapt=1000)
samples <- coda.samples(model, c("theta", "x_pred"),
n.iter=5000)
# Inspecting the posterior
plot(samples)
summary(samples)
# You can extract the mcmc samples as a matrix and compare
# the thetas of the groups. For example, the following shows
# the median and 95% credible interval for the difference
# between Group 1 and Group 2.
samp_mat <- as.matrix(samples)
quantile(samp_mat[, "theta[1]"] - samp_mat[, "theta[2]"], c(0.025, 0.5, 0.975))
```

La línea

```
x[i] ~ dbinom(theta[i], n[i]) JAGS-Code
theta[i] ~ dbeta(1, 1)
```

describe la Likelihood y la Prior asumida. Se puede modificar la Prior si se dispone, por ejemplo, de información contextual que se pueda cuantificar. La Posterior podríamos mirarlo adicionalmente, pero no esperemos nada que no se haya discutido ya:

```
# plot posterior
plot(pres.2x2.bprop)
```

Pasamos a la comprobación de hipótesis a posteriori. Para ello utilizamos el concepto de ROPE (Kruschke, 2015b), definimos un criterio de comparación y calculamos el valor porcentual de la diferencia de θ_1 *y* θ_2, que es menor que dicho criterio. Como criterio elegimos de forma bastante arbitraria $diff_{crit} = 0.03$. Esto da como resultado la probabilidad de que dentro de este intervalo los grupos sean equivalentes (s. cap. 6.8.4.2) o no, más las odds ratio posteriores asociadas.

6.15.3.2 MCMC por fuerza bruta variante 1

La *fuerza bruta* se utiliza comúnmente para describir un enfoque relativamente ciego, a veces muy ineficiente, de la fuerza bruta para llegar a una solución que, sin embargo, es correcta. Un ejemplo son los ataques de piratas informáticos a las contraseñas, en los que, sin comprender demasiado los mecanismos subyacentes, todas las contraseñas posibles son mecanismos, es decir, todas las combinaciones posibles de letras, números y caracteres especiales. Las combinaciones de caracteres especiales se prueban obstinadamente. En la práctica, esto puede llevar mucho tiempo.

Piense en una contraseña larga que contenga letras (grandes y pequeñas), números, caracteres especiales etc. Si elige una contraseña con 45 caracteres y 26 + 26 + 10 + 15 = 77 combinaciones de caracteres diferentes, hay al menos 7745 = 7.799785e + 84 posibilidades de combinación y esto sólo en el caso de que la contraseña tenga *realmente* 45 caracteres (`ptII_quan_Bayes_caso_debates_presidenciales.r`). Si se desconoce la longitud de la contraseña y sólo se sabe que puede tener un máximo de 45 caracteres, pero también menos, se deben probar todos los subconjuntos. también. Esto lleva a un recuento de conjuntos de potencia. Esto supera rápidamente la duración del universo, aunque pudieras probar miles de combinaciones por segundo. Con 100.000 pruebas por segundo, se llega a los citados 45 caracteres se obtiene un teórico $(77^{45})/100000/3600/24/365.25 = 2.471603e+72$ años. El universo tiene actualmente unos 13.81 mil millones de años, es decir, unos 13e+10. Estas ya son diferencias de órdenes de magnitud. Si se utilizaran sólo letras minúsculas y números y la contraseña tuviera sólo 7 caracteres, esto daría como resultado $(26 + 10)^7 = 78\ 364\ 164\ 096$ combinaciones posibles o $(36^7)/100000/3600/24/365:25 = 0.02483211$ años, es decir, algo más de 9 días. Y esto sólo es cierto si la contraseña está bien elegida y da el máximo tiempo mediante fuerza bruta. En la práctica, puede ser más rápido, por ejemplo con ataques de diccionario y contraseñas mal elegidas. La fuerza bruta es, por tanto, el *peor caso* de cómo se puede hacer algo sin pensar lo más lentamente, pero conduce al objetivo.

Para cálculos más pequeños como éste, un enfoque de fuerza bruta es una elegante y legítima variante. La función de R `bayes.prop.mcmc()` toma los valores *a* y *b* de la distribución posterior beta para las distribuciones que se van a probar y el número de simulaciones MCMC `n.mcmc` y repeticiones `nchains`, así como `credMass`, conocido del paquete `BEST` de R para especificar el HDI del intervalo de confianza bayesiano. La función función de R `bprop.mcmc()` también asume una distribución beta(1, 1) como Prior estándar y toma la tabla 2 x 2 de las propiedades. A continuación, llama a la función R `bayes.prop.mcmc()`, que realiza los cálculos reales. Básicamente, el procedimiento es sencillo (`ptII_quan_Bayes_case_presidential-debates.r`):

1. Establecer las distribuciones a priori para ambos grupos. Dado que la distribución prior beta es una distribución conjugada para la Posterior, ésta es también una beta. Por lo tanto, se pueden especificar las Priores a través de los parámetros *a* y *b* de la distribución beta (a_{prior}, b_{prior}). Si no se dispone de información, se puede utilizar una distribución beta uniforme (1, 1) o incluso una Prior de Haldane (Haldane, 1932). Éstas se obtienen seleccionando los parámetros a_{prior} y b_{prior}.
2. Los datos empíricos (= frecuencias) de la 2 x 2 tabla de contingencia siguen como Likelihood un modelo binomial con la lógica de éxito/fracaso o se quiera expresar tal relación binaria. Esto se puede convertir como *s* éxitos de *N* ensayos. La Likelihood con los parámetros *s* y *N* puede transformarse en un valor beta(a_{likeli}, b_{likeli}) según ciertas reglas. Para ello utilizamos la función de R `bino.ab.lik()`.
3. Según las reglas habituales (véase el capítulo 6.12.1) de la distribución beta, las Posteriors se pueden determinar a partir de los parámetros *a* y *b* de las dos Priores y Likelihoods como beta(a_{post}, b_{post}). Esto se hace mediante la función de R `bino.ab.post()`.
4. A partir de los parámetros *a* y *b* de las distribuciones posteriores de los dos grupos, se generan valores a partir de las Posteriores con la función `rbeta()`. Por lo tanto, la prueba real se basa en el siguiente código R abstracto, que genera la cantidad θ_{diff} de interés y comprueba en qué medida es menor que una cantidad de equivalencia `crit.equivalence` que se especificará con más detalle. Todo ello sigue la lógica simple de los muestreadores de Gibbs (s. cap. 6.13.1.2).

```
> # simple gibbs sampler
> seed <- 5555
> set.seed(seed)
> a1 <- 1
> b1 <- 1
> a2 <- 0.6
> b2 <- 5
> n.mcmc <- 1e5
> crit.equivalence <- 0.6
> reps <- 4
> theta.diff <- replicate(reps,
        + rbeta(n.mcmc, shape1=a1, shape2=b1) -
        + rbeta(n.mcmc, shape1=a2, shape2=b2))
> str(theta.diff)
num [1:100000, 1:4] 0.634 0.622 0.886 0.094 0.228 ...
> prob.diff <- apply(theta.diff,2,
        + function(x) mean(x < crit.equivalence))
> prob.diff
[1] 0.70545 0.70569 0.70388 0.70406
```

5. La función de R `replicate()` se utiliza para generar diferentes cadenas de estimaciones MCMC.
6. Se pueden calcular las diferencias de las Posteriores por cadena. La distribución resultante es ya el resultado de la comparación de proporciones.
7. Todo esto se puede evaluar en resumen, las cadenas MCMC se pueden examinar gráficamente y se puede probar la equivalencia de las Posteriores de la diferencia de las dos agrupaciones según un criterio definido o se pueden calcular intervalos de confianza (HDI). El resultado son probabilidades posteriores o Odds Ratios posteriores, que dicen algo sobre la igualdad o desigualdad de los dos grupos.

Pongamos en práctica lo anterior a partir de los datos existentes.

```
# analysis
pres.2x2
pres.2x2.bprop.mcmc <- bprop.mcmc(pres.2x2)
str(pres.2x2.bprop.mcmc)
```

El resultado se puede trazar directamente con `bayes.plot.mcmc()` para comprobar el diagnóstico MCMC (véase la Fig. 6.138).

```
bayes.plot.mcmc(bprop.mcmc.res=pres.2x2.bprop.mcmc)
```

Para examinar la equivalencia de los dos grupos, las hipótesis sobre las diferencias se pueden comprobar del siguiente modo. Si tomamos esto como base, la Figura 6.139 muestra que la probabilidad posterior, es decir la probabilidad de que la diferencia θ_{diff} entre los dos grupos sea superior a 0.14 disminuye significativamente.

Aquí hay algunos valores como ejemplos:

```
> # diagnostics
> # posterior Odds Ratios
> credMass <- 0.99
> mean(abs(pres.2x2.bprop.mcmc$mcmc[["theta (diff)"]]) < (1-credMass))
[1] 0.135
> # better ROPE
> mean(abs(pres.2x2.bprop.mcmc$mcmc[["theta (diff)"]]) < 0.03)
[1] 0.388
> mean(abs(pres.2x2.bprop.mcmc$mcmc[["theta (diff)"]]) < 0.05)
[1] 0.59425
> mean(abs(pres.2x2.bprop.mcmc$mcmc[["theta (diff)"]]) < 0.1)
[1] 0.917
> #etc... further hypotheses possible
```

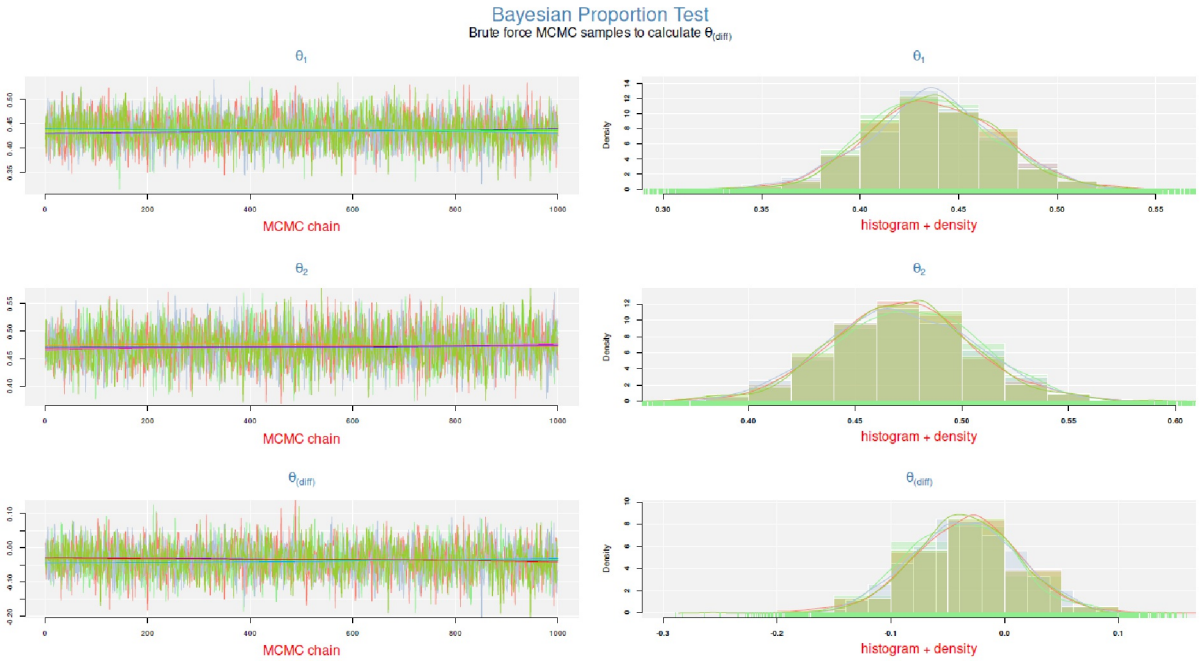


Figura 6.138. *I, we y nation*
(prueba de proporción de Bayes mediante fuerza bruta, variante 1, diagnóstico MCMC)

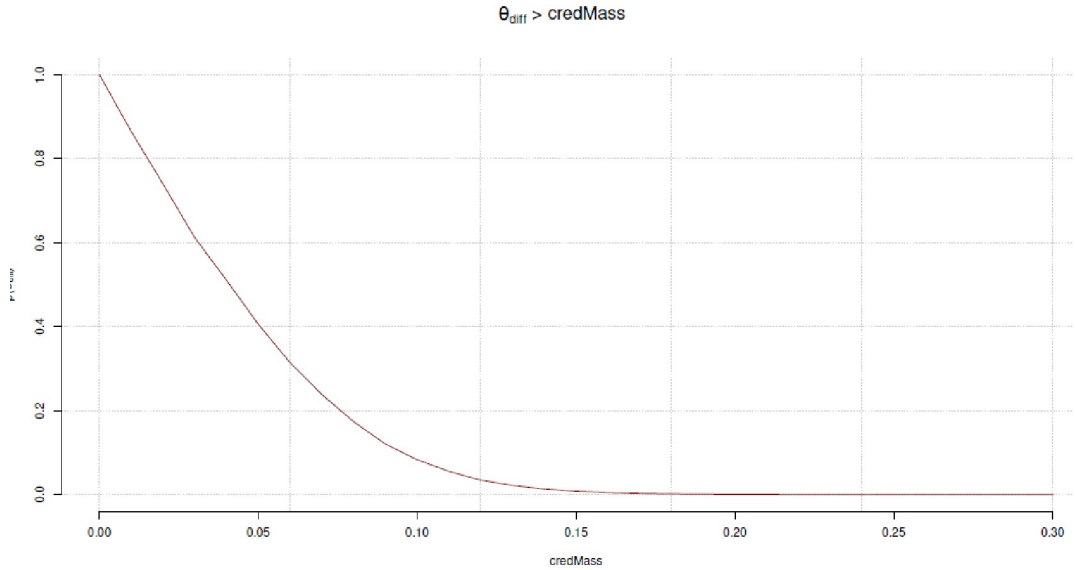


Figura 6.139. *I, we y nation*
(prueba de proporción de Bayes mediante fuerza bruta, variante 1, hipótesis θ_{diff} > credMass)

Se pueden trazar muchas pruebas de hipótesis sucesivas para comprender mejor el área de interés. Esto no es más que un ROPE alrededor de un área de interés para nosotros:

```
# plot sequence of hypotheses R-Code
# define credMass area
sek <- seq(0,0.3,0.01)
bprop.vs.sek <- sapply(sek,
        function(i)
        {
        mean(abs(pres.2x2.bprop.mcmc$mcmc[["theta (diff)"]]) > i))
        }
bprop.vs.sek
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l")
plot(sek, bprop.vs.sek, type="l", bty="n",
        col="darkred", pre.plot=grid(),
        xlab="credMass",
        ylab=eval(substitute(expression(paste("p(",theta[diff],")",
        sep="")))), main="")
mtext(eval(substitute(expression(paste(theta[diff]," > credMass",
        sep="")))), outer=TRUE, line=-2, cex=1.5, side=3)
```

Ahora extraemos los parámetros posteriores a_1, b_1 y a_2, b_2 de las distribuciones beta para utilizarlos posteriormente en caso de que sea necesario.

```
> pres.2x2.bprop.mcmc <- bprop.mcmc(pres.2x2)
> pres.post.betavalues <- pres.2x2.bprop.mcmc$a1b1a2b2
> names(pres.post.betavalues) <- c("a1","b1","a2","b2")
> pres.post.betavalues
a1  b1  a2  b2
102 132 108 121
```

6.15.3.3 MCMC mediante fuerza bruta variante 2

El script de R `ptII_quan_Bayes_case-prestermusage.r` contiene una segunda variante que sigue la misma lógica interna que *la variante de fuerza bruta 1*, pero utiliza una lista y `lapply()` en lugar de `replicate()`. Esto es sólo para mostrar que R ofrece formas bastante diferentes de implementar la misma cosa de forma flexible. Por lo tanto, sólo imprimimos el código R en extractos. Los detalles se pueden encontrar en los scripts de R (`ptII_quan_Bayes_case_presidential-debates.r`).

En primer lugar, se definen el número de cadenas MCMC, el número de extracciones MCMC a partir de los valores posteriores y la lista de resultados.

```
# add chains
nchains <- 3
n.mcmc <- 1e+4
theta.diff.mcmc <- gr1.mcmc <- gr2.mcmc <- vector(mode="list", length=nchains)
```

Tomamos los valores de las distribuciones beta del vector `pres.post.betavalores` que acabamos de crear anteriormente. Esto es seguido por la llamada con `lapply()` para generar las Posteriores.

```
# calculate x chains of MCMC for group1 and group2
set.seed(78824)
a1b1a2b2 <- pres.post.betavalues
# posteriors
gr1.mcmc <- lapply(gr1.mcmc,
        function(x) x <- rbeta(n.mcmc, shape1=a1b1a2b2[1], shape2=a1b1a2b2[2]))
gr2.mcmc <- lapply(gr2.mcmc,
        function(x) x <- rbeta(n.mcmc, shape1=a1b1a2b2[3], shape2=a1b1a2b2[4]))
```

...y se utiliza un simple bucle `for()` para generar la cantidad θ_{diff} de interés.

```
# calculate theta.diff = theta1 - theta2
for(i in 1:nchains)
{
theta.diff.mcmc[[i]] <- gr1.mcmc[[i]] - gr2.mcmc[[i]]
}
```

Todo lo demás son llamadas a listas para los estadísticos de resumen, HDI, etc. a través de las cadenas MCMC más las comprobaciones gráficas habituales (véase la Fig. 6.140). Omitimos este código R aquí. La prueba real se parece a la variante 1. Usamos la lista generada `theta.diff.mcmc`, que contiene la diferencia θ_{diff} de interés por cadena MCMC.

```
# test some hypotheses R-Code
# Probability theta1 > theta2 and theta1 < theta2
# over all chains
critv <- 0
# remember: diff = theta1 - theta2
# diff > 0 -> theta1 > theta2
diffGREATERzero <- sapply(theta.diff.mcmc, function(x) mean(x > critv))
# diff < 0 -> theta1 < theta2
diffSMALLERzero <- sapply(theta.diff.mcmc, function(x) mean(x < critv))
```

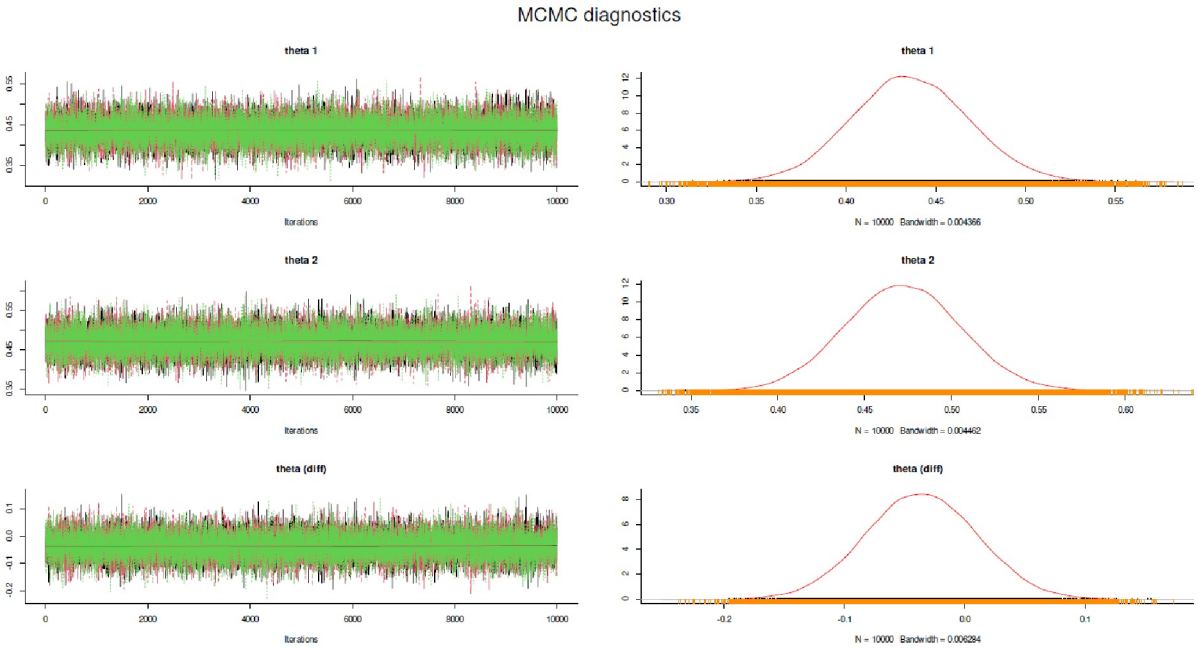


Figura 6.140. *I, we y nation*
(prueba de proporción de Bayes mediante fuerza bruta, variante 2, diagnósticos MCMC).

Calculamos así una odds ratio posterior a favor de $\theta_1 > \theta_2$ y observamos previamente la salida de las hipótesis (cada una separada por el MCMC):

```
> # output of hypotheses
> diffGREATERzero
[1] 0.2155 0.2240 0.2192
> diffSMALLERzero
[1] 0.7845 0.7760 0.7808
>
> # posterior ratio
> diffGREATERzero / (1-diffGREATERzero)
[1] 0.2746973 0.2886598 0.2807377
```

y vice versa:

```
> # posterior odds
> diffSMALLERzero / (1-diffSMALLERzero)
[1] 3.640371 3.464286 3.562044
```

La probabilidad de que $\theta_1 - \theta_2 > 0$ es $p = 0.216$ y se mantiene $p = 0.785$ que la masa de la distribución de Kerry es típicamente mayor que la de Bush. Pero si preguntamos la magnitud, esta tesis general se derrumba a

```
> critv <- 0.1
> sapply(theta.diff.mcmc, function(x) mean(x > critv))
[1] 0.0022 0.0021 0.0016
```

de modo que una $p = 0.0022$ significa que la diferencia es superior a 0.1. Esto demuestra una vez más que las afirmaciones generales sólo son fiables hasta cierto punto, pero que la cuestión de la dirección y la magnitud de los efectos siempre debe plantearse a continuación. Las probabilidades posteriores pueden convertirse en Odds Ratios como se muestra arriba – al final, no sale nada, salvo que la información se presenta de forma diferente. Uno puede entonces decidir por sí mismo si está mejor con las probabilidades o con las proporciones.

También podría interesarnos la probabilidad posterior de que la desviación absoluta y no dirigida comparada con la desviación dirigida esté por encima o por debajo de un determinado criterio `critv`. Repetimos el análisis anterior para los valores absolutos de las diferencias:

```
> # ask about absolute difference without any direction
> sapply(theta.diff.mcmc, function(x) mean(abs(x) > critv))
[1] 0.0888 0.0842 0.0897
```

Un vistazo a la visualización de di o el valor absoluto de di visualiza la diferencia con y sin valores absolutos (véase la Fig. 6.141, código R no impreso).

Recordatorio 6.5: ¡Piensa de otra manera!

Pensemos que en lugar de preguntar por la diferencia entre las alturass corporales, preguntamos por su la igualdad ¿Qué esperamos? ¿Identidad? En absoluto, porque sabemos que cuanto más exactamente medimos, más diferencias se ponen de manifiesto y difícilmente encontraremos dos aspirantes *y* presidentes que tuvieran alguna vez el 100% de la misma talla.

Es mucho más importante cómo de grande puede cuantificarse una diferencia y si hay algún tipo de significación asociada a ella. ¿Están los votantes básicamente cegados por el hecho de que siempre votan al candidato más grande e ignoran lo que dicen los demás? Si es así, las mujeres tendrían muchas menos posibilidades, ya que por término medio son algo más pequeñas que los hombres, pero esto no cubre muchas constelaciones en las que no es así.

Entonces, ¿qué preguntas queremos hacer a los datos? ¿Sólo las relativas a la diferencia? En ya muestra que esto no es muy fructífero. En este punto se hace evidente que la inclusión del concepto ROPE (véase el capítulo 6.8.4.2) como consideración básica para introducir rangos de tolerancia antes de hablar de significación debería ser una sabia decisión en la mayoría de los casos.

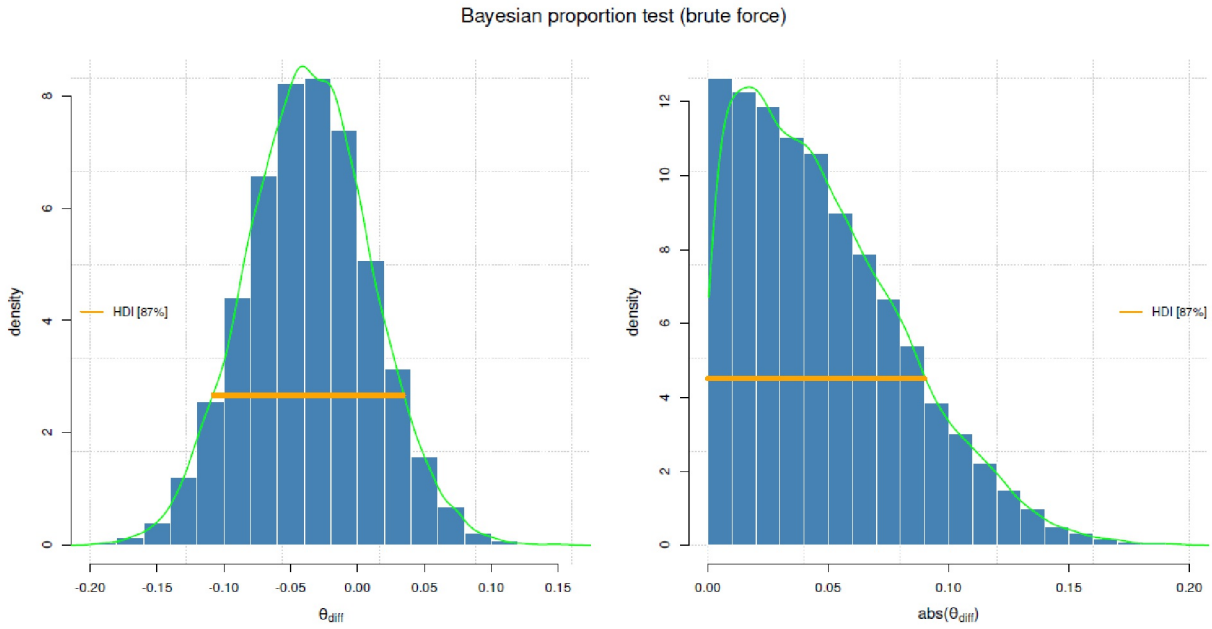


Figura 6.141. *I, we y nation*
(prueba de proporción de Bayes por fuerza bruta, variante 2, θ_i posterior y sus valores absolutos)

6.15.3.4 Integración numérica o aproximación mediante un cadrículo

La aproximación mediante un cadrículo corresponde a la idea de utilizar una secuencia de valores que son válidos como valores iniciales de una función para obtener los correspondientes valores de la función exactamente en esos valores. Si el espaciado se elige lo suficientemente estrecho, se obtiene una aproximación bastante buena a la función continua real. Esto significa que una función continua se utiliza realmente de forma discreta y la divide en trozos calculables. La secuencia se puede elegir de esta manera – en el caso de la distribución beta utilizada aquí mediante `dbeta()` para calcular la densidad de la distribución beta, el rango de valores es generalmente de 0 a 1, ya que este es el rango de valores de la probabilidad. Sin embargo, el procedimiento es bastante ineficiente, pero funciona muy bien en la práctica.

Para una mejor comprensión imprimimos completamente la función R `bayes.prop.grid()` (véase en `ptII_quan_Bayes_caso_debates_presidenciales.r`).

```
bayes.prop.grid <- function(a1=a1, b1=b1, a2=a2, b2=b2, int.width=1e-3,
       start.sek=0, end.sek=1)
{
sek <- seq(start.sek, end.sek, int.width)
# important part: (x > y)
grid.res <- outer(sek, sek,
       function(x, y) (x > y) * dbeta(x, a1, b1)
       *int.width * dbeta(y, a2, b2) * int.width)
# (y - x <0.1)
# grid.res <- outer(sek, sek,
       # function(x, y) (y - x <0.1) *
           dbeta(x, a1, b1)*int.width *
           dbeta(y, a2, b2)*int.width)
return(grid.res)
}
```

En primer lugar, la secuencia `sec` se define con la anchura de intervalo `int.width`, de modo que si `int.width` = 1e-3 = 0.001 hay exactamente 1/`int.width` = 1/0.001 = 1 000 elementos de cuadrícula. Para los valores de la cuadrícula, necesitamos la función de R `outer()`, que genera el producto exterior de dos vectores. La forma en que funciona se hace evidente cuando generamos el *uno a uno* de esta manera:

```
> # outer product
> 1:10 %o% 1:10
[,1]  [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]   1    2    3    4    5    6    7    8    9   10
[2,]   2    4    6    8   10   12   14   16   18   20
[3,]   3    6    9   12   15   18   21   24   27   30
[4,]   4    8   12   16   20   24   28   32   36   40
[5,]   5   10   15   20   25   30   35   40   45   50
[6,]   6   12   18   24   30   36   42   48   54   60
[7,]   7   14   21   28   35   42   49   56   63   70
[8,]   8   16   24   32   40   48   56   64   72   80
[9,]   9   18   27   36   45   54   63   72   81   90
[10,] 10   20   30   40   50   60   70   80   90  100
```

que es idéntico a

```
outer(1:10,1:10,"*")
```

Para cada elemento de la cuadrícula, se calcula el producto de las respectivas densidades beta multiplicado por la anchura del intervalo (= escalado) y se combina con la consulta de si $\theta_1 > \theta_2$. En lugar de la pregunta $\theta_1 > \theta_2$, se podría formular cualquier hipótesis que tuviera una respuesta razonable, por ejemplo una basada en equivalencia como $\theta_1 - \theta_2 < \text{diff}_{\text{crit}}$. Las probabilidades a posteriori o los Odds Ratios a posteriori resultan de las sumas de la cuadrícula a lo largo de la dirección de la hipótesis, específicamente:

```
> # calculate based on grid approximation
> prob.a1b1.vs.a2b2 <- bayes.prop.grid(a1=a1, b1=b1,
a2=a2, b2=b2,
int.width=1e-3)
> # prob theta1 > theta2
> sum.prob <- sum(prob.a1b1.vs.a2b2)
> sum.prob
[1] 0.2164488
> # prob theta1 < theta2
> 1 - sum.prob
[1] 0.7835512
> # odds ratios
> sum.prob / (1-sum.prob)
[1] 0.2762408
> (1-sum.prob) / sum.prob
[1] 3.62003
```

Si observamos los resultados para la beta(102, 132) y la beta(108, 121) posteriores, respectivamente, no parece una distribución desigual. La hipótesis $\theta_1 > \theta_2$ tiene una probabilidad posterior de $p_{gridapprox} = 0.216$ y una Odds Ratio posterior de $OR_{gridapprox} = 0.276$. No son valores sobresalientes que puedan invalidar los análisis anteriores. Además los valores son muy similares a las explicaciones anteriores.

6.15.3.5 Solución analítica exacta (según Evan Miller, 2015)

La prueba de proporción bayesiana conoce una solución analíticamente exacta – como la prueba *t* baye-siana (Bretthorst, 1993). Miller (2015) la deriva analíticamente (cf. Raineri, Dabad & Heath, 2014-05-13). La función R resultante `h()` prueba $\theta_2 > \theta_1$ (`ptII_quan_Bayes_case_presidential-debates.r`).

```
a1 <- pres.post.betavalues["a1"]
b1 <- pres.post.betavalues["b1"]
a2 <- pres.post.betavalues["a2"]
b2 <- pres.post.betavalues["b2"]
# important: if loga=TRUE
# result is a BROB object
# Pr(GR2 > GR1)
h.res <- h(a1=a1, b1=b1, a2=a2, b2=b2)
h.res
# log version
h.res.log <- h(a1=a1, b1=b1, a2=a2, b2=b2, loga=TRUE)
h.res.log
as.numeric(h.res.log)
```

Esto cae con $p_{exact} = 0.2196$ bastante cerca de la solución de aproximación de cuadrículo anterior. Lo mismo ocurre con La Odds Ratio posterior y las demás probabilidades:

```
> # Probability(GR2 > GR1)
> h.res
[1] 0.7803622
> # Pr(GR2 < GR1)
> 1 - h.res
[1] 0.2196378
> # Ratio in favor of GR2 > GR1
> h.res / (1-h.res)
[1] 3.552951
# GR2 < GR1
> 1/( h.res / (1-h.res) )
[1] 0.2814562
> # Odds ratio in favor of GR2 > GR1
> ((h.res)/(1-h.res)) / ((1-h.res)/h.res)
[1] 12.62346
> # GR2 < GR1
> ((1-h.res)/h.res) / ((h.res)/(1-h.res))
[1] 0.07921757
```

El resultado es una $Ratio_{exact} = 0.281$ que también se aproxima mucho a la solución anterior mediante cuadrícula. Las diferencias de los enfoques se pueden cuantificar, tomamos la hipótesis anterior `sum.prob`:

```
> # difference between brute force numerical integration and h()
> h.res.inv <- 1-h.res
> # difference
> sum.prob - h.res.inv
[1] -0.003161649
> # equal
> 1 - abs(sum.prob - h.res.inv)
[1] 0.9968384
> # ratio
> sum.prob / h.res.inv
[1] 0.9856052
```

A la vista de este alto grado de coincidencia, deberíamos mostrar una confianza creciente en los planteamientos bayesianos que llegan a soluciones prácticamente idénticas de formas bastante diferentes. Lo mismo se aplica a la solución con el paquete R `BayesianFirstAid` (véase más arriba, resultados no impresos).

Stucchio (2014b, 2014a, 2015), basándose en el mencionado trabajo de Miller, amplía este enfoque y añade una función de pérdida para refinar las pruebas A/B bayesianas (= comparación de dos tasas de éxito, es decir, dos proporciones) y las reglas de decisión asociadas. Siguiendo las explicaciones del autor, la función de R `bayes-prop.loss()` hace esto a pequeña escala. La función espera los valores beta de las dos proporciones, así como un umbral crítico, de modo que: $\theta_2 - > \theta_1 <$ crit. [=prueba]. Especificamos el criterio como 1- `credMass`, el procedimiento elegido para los HDI. La salida contiene los valores de pérdida respectivos para cada grupo, su diferencia más la prueba, independientemente de que esta diferencia se mantenga o no por debajo del valor crítico. Además indica la columna `LOG` si los datos se procesaron en la escala logarítmica (que es la predeterminada).

```
> # [theta2 - theta1] < crit (Test)
> credMass <- 0.99
> res <- bayes.prop.loss(a1=a1, b1=b1,
        + a2=a2, b2=b2, crit=1-credMass)
### Bayesian A/B Testing ###
Test [Group_2 - Group_1] < crit
a1 = 102 , b1 = 108
vs.
a2 = 132 , b2 = 121
  LOG  loss GR1   loss GR2  loss [GR2-GR1] crit
1 TRUE 0.04272179 0.1417564 0.1086484      0.01
credMass loss [GR2-GR1] < crit
    1      0.99           FALSE
> # change criteria
> credMass <- 0.1
> res <- bayes.prop.loss(a1=a1, b1=b1,
        + a2=a2, b2=b2, crit=1-credMass)
### Bayesian A/B Testing ###
Test [Group_2 - Group_1] < crit
a1 = 102 , b1 = 108
vs.
a2 = 132 , b2 = 121
  LOG  loss GR1   loss GR2  loss [GR2-GR1] crit
1 TRUE 0.04272179 0.1417564 0.1086484      0.9
credMass loss [GR2-GR1] < crit
    1      0.1            TRUE
```

Ahora queremos saber dónde cambia el valor de FALSE a TRUE:

```
> # table TRUE vs. FALSE
> loss.v <- res[,4]
> sek <- seq(0,1,0.001)
> tab <- table(loss.v < sek)
> tab/sum(tab)
FALSE TRUE
0.1088911 0.8911089
> # when does it change from FALSE to TRUE
> tf.IDs <- which(( lossvBELOWsek <- loss.v < sek) == TRUE)[1]
> data.frame(loss.v,sek, "loss.v < sek"=lossvBELOWsek,
        check.names=FALSE)[(tf.IDs[1]-1):tf.IDs[1],]
    loss.v sek loss.v < sek
109 0.1086484  0.108    FALSE
110 0.1086484  0.109    TRUE
```

6.15.3.6 Solución analítica exacta

(según Pham-Gia, Turkkan y Eng (1993 resp. Nadarajah y Kotz, 2007).

Otra solución analítica exacta se aplica en Sverdlov, Ryeznik y Wu (2015) utilizando código R y se basa en los trabajos de T. Pham-Gia, N. Turkkan y P. Eng (1993) y Nadarajah y Kotz (2007) y en las correcciones de Chen y Luo (2011). Hemos reprogramado y modificado ligeramente las funciones R de los autores. Dado que ahora estas utilizan funciones R compiladas del paquete R `appell` y este se eliminará en 2019de CRAN debido a la falta de mantenimiento del paquete, solo las enumeramos en extractos (ptII_quan_-Bayes_case_presidential-debates.r).. Sin embargo, ciertamente se puede cargar la biblioteca compilada de una versión anterior de R con

```
dyn.load("appell/libs/i386/appell.dll")
```

Atención: ¡adapte primero la ruta anterior a las condiciones locales! y luego copie manualmente los scripts R `appellf1()` o `hyp2f1()` de `appell` a R o intégrelos con `source()`. En *Windows 7* con *R 3.6.0* esto nos funcionó sin problemas. En *Linux* no funciona; no lo hemos probado con *Wine*. Esto probablemente funcionaría siempre y cuando la inclusión del código compilado no cambie en R o cuando podamos enlazar todas las librerías necesarias estáticamente. El paquete R que utilizamos, `appell`, fue compilado bajo *R 3.4.0*. Además puesto que aquí no se propaga ningún espacio de nombres en nuestra solución, las subrutinas de las llamadas `Fortran()` deben entrecomillarse. En forma abreviada esto se ve como `hyp2f1()`

```
# replace f21_sub by „f21_sub“ (with quotation marks)
results <- .Fortran("f21_sub", a = a, b = b, c = c, z = z,
       hyp2f1 = algorithm, val = val)
```

y en `appellf1()` como

```
# replace f1 by „f1“ (with quotation marks)
results <- .Fortran("f1", a = a, b1 = b1, b2 = b2, c = c, x = x,
       y = y, algoflag = algoflag, userflag = userflag,
       debug = debug,
       val = val, hyp2f1 = hyp2f1)
```

En Linux, se puede obtener las fuentes en Internet y compilarlas por sí mismo y, si es necesario, integrarlas manualmente como se ha descrito anteriormente. Mientras tanto, existe el paquete R `tolerance`, que proporciona las funciones `F1()` de R con la función hipergeométrica `F1` de `Appell`. Además, existen `ddiffprop()`, `pdfiffprop()`, `qdiffprop()` y `rdiffprop()`, que producen las funciones de densidad, probabilidad, cuantil y aleatoria para la diferencia de dos proporciones.

Las funciones también se basan en el trabajo de Nadarajah y Kotz (2007) y las modificaciones de Chen y Luo (2011). En el paquete de R `hypergeo` existe `hypergeo()`, que sustituye a la función de R `hyp2f1()` y que utilizamos internamente en lo que sigue. Entonces todo funciona sin problemas tanto en Windows como en Linux con las versiones actuales de R. Por lo tanto, la solución anterior no es necesaria, pero puede ser útil como idea en otros casos cuando se trata de integrar funciones compiladas.

A continuación, la función de R `prop.theta.sec()` y sus subrutinas se han reelaborado de nuevo para que ahora también puedan funcionar sobre la base de logaritmos mediante el paquete de R `Brobdingnag` si los valores se vuelven demasiado grandes o demasiado pequeños entretanto y abandonan temporalmente el espacio de la resolución numérica normal en el ordenador. Al utilizar esta prueba exacta, se observa que tiene inestabilidades numéricas en el rango alrededor de 0. Estas fueron confirmadas por el autor Sverdlov vía correo electrónico. Una solución per se no es aparente,pero el problema parece ocurrir especialmente cuando dos proporciones difieren muy poco. Aparte de estas inestabilidades numéricas, sin embargo, esta

prueba funciona exactamente, lo que examinamos con más detalle en comparación con la variante de fuerza bruta.

La función de R `prop.theta.sec()` basada en el trabajo de Sverdlov, Ryeznik y Wu (2015) y su código de R toma los valores posteriores de la distribución beta como de costumbre, así como criterios para la diferencia de los valores θ y de comparación para la proporción Relative Risk θ_1/θ_2 y la Odds Ratio. Las variables `xlim.diff, xlim.RR` y `xlim.OR` dan los límites inferior y superior de la función de densidad de probabilidad (= PDF) o función de densidad acumulativa (= CDF) para diferencias $\theta_1 - \theta_2$, Ratio θ_2/θ_1 y Odds Ratio $(\theta_2 / (1-\theta_2)) / ((\theta_1 / (1-\theta_1))$, mientras que `theta.crit` es un simple criterio de comparación para el gráfico posterior e instruye a `loga` con valores de verdad cuál de los tres valores – diferencias, proporciones u odds ratios – debe calcular en la escala `log()` para evitar mensajes de error debidos a números demasiado grandes. Con `numer` se pueden mostrar los valores originales de las integrales para fines de depuración, o se pueden acotar los límites de las integrales para un examen más detallado con `sL` y `sH`. Con `parallel=TRUE`, la subrutina puede utilizar más de un núcleo de CPU para números grandes.

En primer lugar, comparamos los datos, tomados del script R `bayesian2beta.r` (que está disponible como apéndice del artículo de Sverdlov, Ryeznik y Wu (2015)) con la variante de Fuerza Bruta (véanse Fig. 6.142 y Fig. 6.143).

```
# values from paper from Sverdlov, Ryeznik and Wu 2015 R-Code
a1 <- 1/3+7
b1 <- 1/3+12-7
a2 <- 1/3+6
b2 <- 1/3+18-6
theta.res <- prop.theta.sek(a1=a1, b1=b1, a2=a2, b2=b2,
       loga=c(T,T,T),
       parallel=TRUE,
       numer=TRUE,
       BROB=c(T,T,T),
       xlim.diff=c(-1,1,-1,1), l.diff=100,
       xlim.RR=c(0,10, 0,2), l.RR=100,
       xlim.OR=c(0,100, 0,2.5), l.OR=100,
       theta.crit=c(0.5,1,1)
```

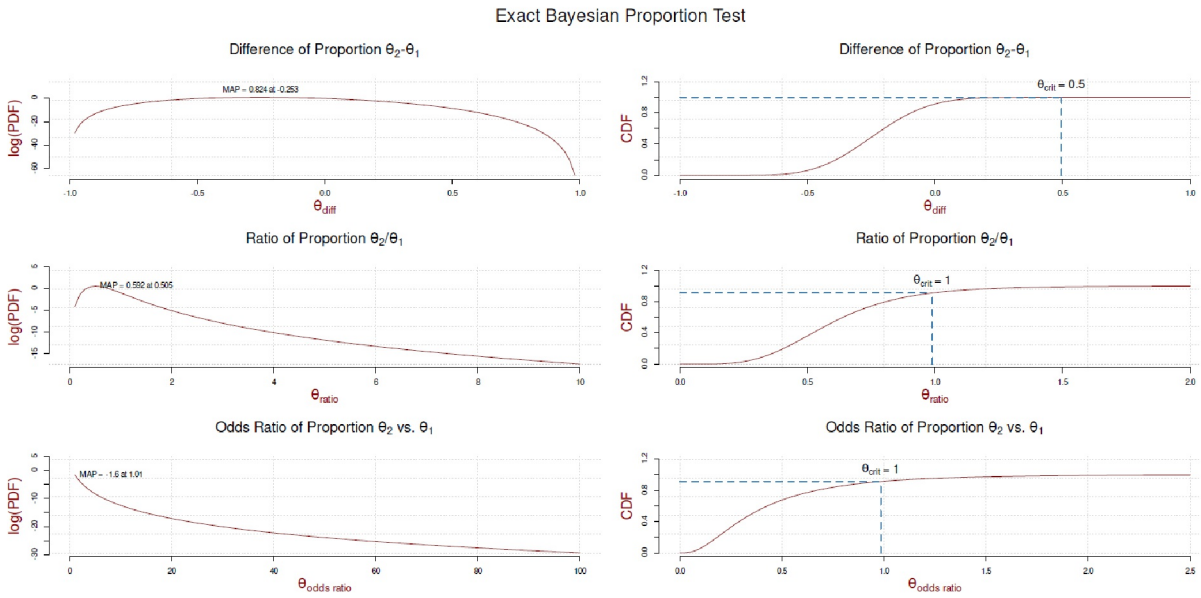


Figura 6.142. *Prueba de proporción de Bayes mediante prueba exacta (θ_{diff}, RR y OR, datos de Sverdlov, Ryeznik & Wu, 2015)*

Y ahora el enfoque de Fuerza Bruta:

```
# brute force
set.seed(192934)
bf.N <- 10e6
bf.res1 <- rbeta(n=bf.N, shape1=a1, shape2=b1)
bf.res2 <- rbeta(n=bf.N, shape1=a2, shape2=b2)
bf.diff <- bf.res2 - bf.res1
plotPost(bf.diff, credMass=0.87, ROPE=c(-0.5,0.5),
        xlab="theta difference", showMode=TRUE,
        col="skyblue", border="white", compVal=0.5)
```

También con los MAPs

```
# MAP exact
theta.res.exp <- theta.res$post$differ$pdf
theta.res.exp[,2] <- exp(theta.res.exp[,2])
MAP.xct <- theta.res.exp[(theta.res.exp[,2] ==
        max(theta.res.exp[,2])),]
# MAP brute force
MAP.bf <- mean(bf.diff)
```

Sigue la salida de los valores máximos de las Posteriores:

```
> # output and comparison
> MAP.xct
   sek.pdf.diff pdf.diff
38 -0.2525253   2.279512
> MAP.bf
[1] -0.2395774
> (MAP.xct[,1]-MAP.bf)/MAP.xct[,1]
[1] 0.0512734
```

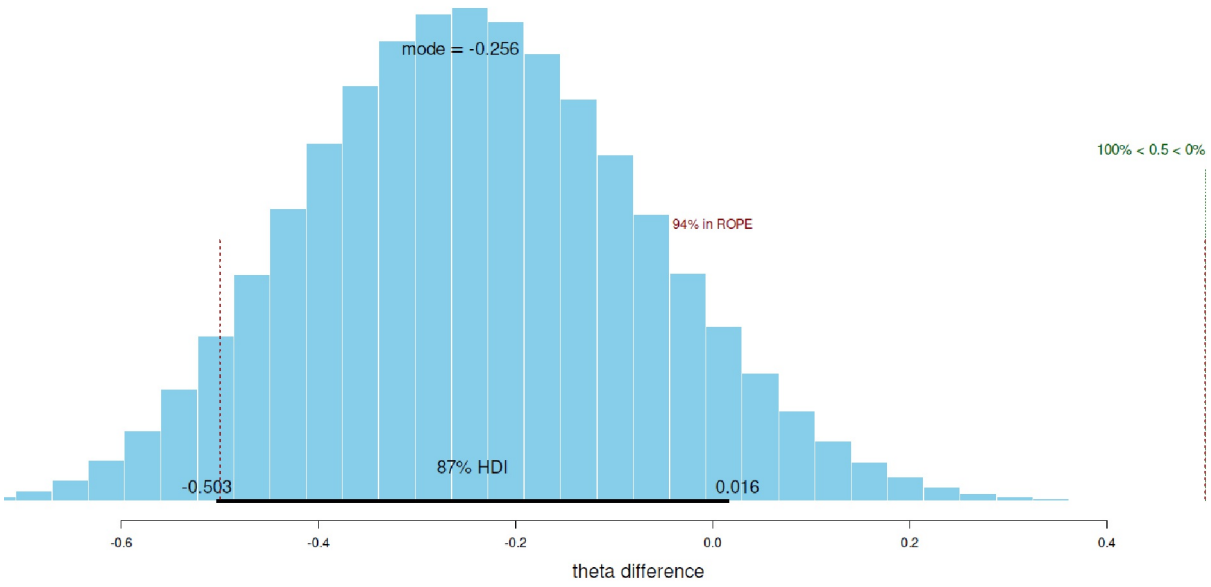


Figura 6.143. *Prueba de proporción de Bayes mediante fuerza bruta* (θ_{diff}, *datos de Sverdlov, Ryeznik yWu, 2015*).

Como se puede ver, los resultados apenas difieren apreciablemente. Lo mismo muestran los gráficos, una vez sin y con logaritmación (véanse Fig. 6.144, izquierda y derecha respectivamente). Los valores coinciden entre la prueba exacta y el enfoque de fuerza bruta.

```
## plot
bf.diff.dens <- density(bf.diff)
par(mfrow=c(1,2))
# non-log
plot(theta.res.exp[,1], theta.res.exp[,2], type="l", col="darkred",
       bty="n", pre.plot=grid(), xlab="theta diff", ylab="pdf.diff")
abline(v=MAP.xct[,1], col="darkred")
lines(bf.diff.dens$x, bf.diff.dens$y, col="blue")
abline(v=MAP.bf, lty=3, col="blue")
# log
theta.res.log <- theta.res$post$differ$pdf
plot(theta.res.log[,1], theta.res.log[,2], type="l", col="darkred",
       bty="n", pre.plot=grid(), xlab="theta.diff", ylab="log(pdf.diff)")
abline(v=MAP.xct[,1], col="darkred")
lines(bf.diff.dens$x, log(bf.diff.dens$y), col="blue", lty=3, lwd=2)
abline(v=MAP.bf, lty=3, col="blue")
mtext(expression(
       paste("Bayesian Analysis of Difference of Proportions ",theta[2],
       " - ",theta[1]," | Exact vs. Brute Force",sep="")),
       3, line=-2, cex=1.6, outer=TRUE)
```

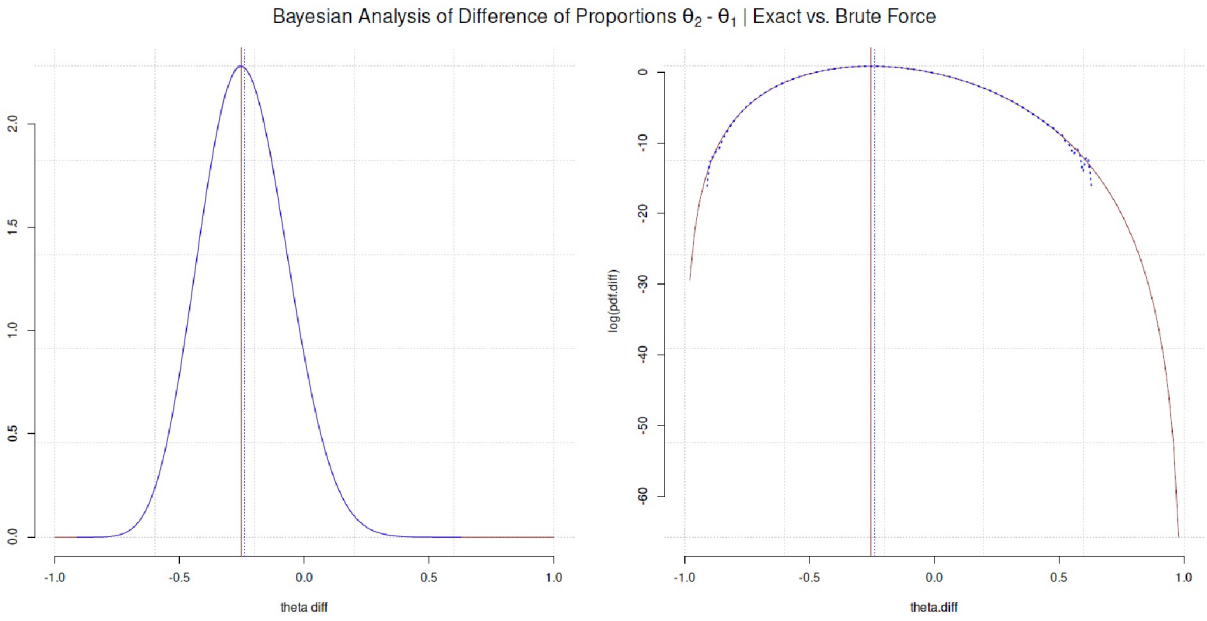


Figura 6.144. *Prueba de proporción de Bayes*
(comparación de θ_{diff}, prueba exacta frente a fuerza bruta, datos de Sverdlov, Ryeznik, & Wu, 2015)

A continuación, pasamos a los datos presidenciales, que ahora también incluyen la comparación de los Ratios u Odds Ratios de las proporciones. Recordemos los datos posteriores:

```
# Bush vs. Kerry
> pres.post.betavalues
a1  b1  a2  b2
102 108 132 121
```

Aquí siguen los cálculos de diferencia, Ratio y Odds Ratio (véase la Fig. 6.145):

```
a1 <- pres.post.betavalues["a1"] R-Code
b1 <- pres.post.betavalues["b1"]
a2 <- pres.post.betavalues["a2"]
b2 <- pres.post.betavalues["b2"]
theta.res <- prop.theta.sek(a1=a1, b1=b1, a2=a2, b2=b2,
       loga=c(T,T,T),
       parallel=TRUE,
       numer=TRUE,
       BROB=c(T,T,T),
       xlim.diff=c(-1,1,-1,1), l.diff=100,
       xlim.RR=c(0,10, 0,2), l.RR=100,
       xlim.OR=c(0,100, 0,2.5), l.OR=100,
       theta.crit=c(0.5,1,1)
       )
```

La función de R `prop.theta.sec()` también produce gráficos por defecto (ver Fig. 6.145) con `prop.theta.sec.plot()` para la PDF y la CDF y por separado para la diferencia de proporciones $\theta_2 - \theta_1$, los Ratios θ_2/θ_1 y la Odds Ratio θ_2 frente a θ_1, es decir, $OR_{(\theta 2 \text{ vs.} \theta 1)} = (\theta_2 /(1-\theta_2))/ (\theta_1/(1-\theta_1))$. Individualmente, el gráfico se puede llamar con el R código

```
prop.theta.sec.plot(theta.res)
```

Con valores mayores, parece diferente del ejemplo anterior del artículo de Sverdlov, Ryeznik y Wu, 2015. Aplicando el procedimiento a los datos presidenciales y a la diferencia de las proporciones como prueba exacta (véase la Fig. 6.145) o como fuerza bruta (véase la Fig. 6.146 e conjuntamente, véase la Fig. 6.147), código R no impreso.

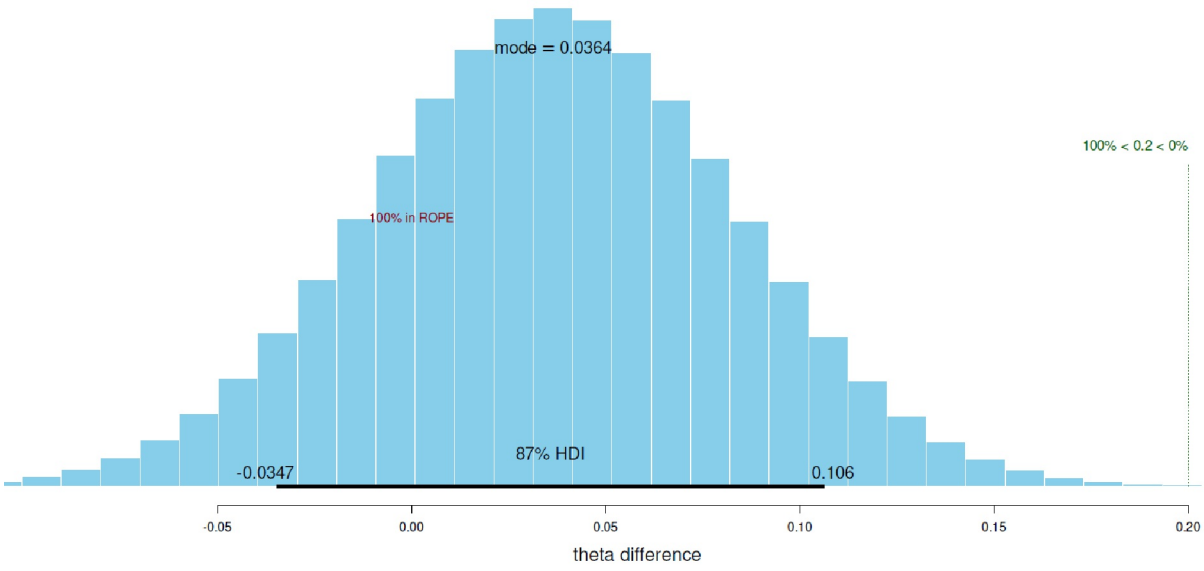


Figura 6.146. *I, we y nation* (θ_{diff}, *prueba de proporciones baysiana mediante Fuerza Bruta)*

Los MAP muestran aquí diferencias:

```
> # output and comparison
> MAP.xct
sek.pdf.diff pdf.diff
51 0.01010101 9.12595
> MAP.bf
[1] 0.03600788
> (MAP.xct[,1]-MAP.bf)/MAP.xct[,1]
[1] -2.56478
```

En nuestra opinión, se encuentra la causa en las inestabilidades numéricas descritas, es decir, en el cálculo de la integral de la distribución hipergeométrica en el intervalo de 0 y 1. Una hipótesis de trabajo sería que el problema surge sobre todo cuando hay un mayor número de casos y sólo una diferencia empírica muy pequeña, como ocurre en este caso. Esto podría explicarse por la Figura 6.147. En ella se compara la prueba exacta con el método de fuerza bruta. Aquí el gráfico de la izquierda muestra que el máximo de la prueba exacta está muy cerca de 0. Más concretamente, se debe al uso de la distribución hipergeométrica, que en los extremos de la integral en 0 y 1, cuando se examina la diferencia de las proporciones alrededor de la zona 0 y la diferencia exacta sólo se desvía ligeramente aquí. Esto tiene un efecto distorsionador en los cálculos posteriores. En ausencia de estos datos extremos, la prueba exacta y el enfoque de fuerza bruta coinciden prácticamente, como muestra el ejemplo anterior de los datos de Sverdlov, Ryeznik y Wu (2015) con bastante claridad. Se podrían observar los valores correspondientes de las integrales respectivas por separado para comprobar y comprender mejor el efecto de los valores extremos (véase también para comprender mejor el efecto de los valores extremos (véanse las Fig. 6.148 y 6.149). El enfoque de fuerza bruta es una necesidad válida para comprobar la plausibilidad de los resultados.

Veamos dicha integral para los valores de 0 a 1 sobre todos los valores posibles de θ_{diff} (R-código no impreso, ver Fig. 6.148 para un gráfico 2D y Fig. 6.149 para un gráfico 3D, los valores infinitos fueron eliminados para el gráfico). El gráfico 3D se generó con `persp3D()` del paquete R `plot3D`. Para una mejor legibilidad, se puede limitar el rango *Y* trazando sólo sobre ciertos cuantiles de la distribución.

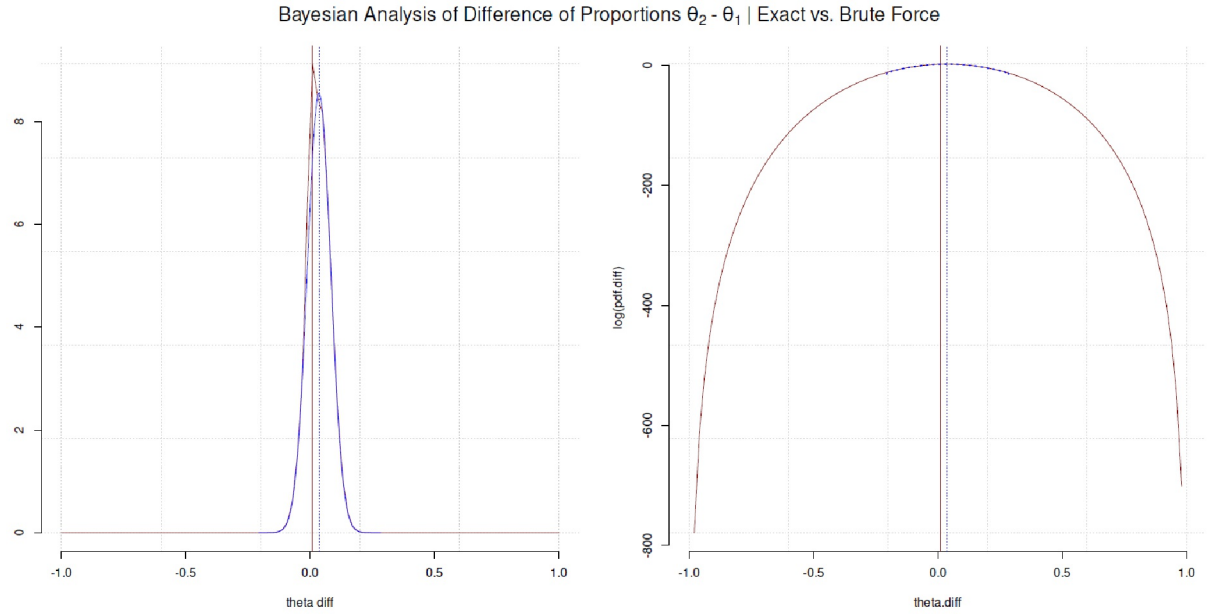


Figura 6.147. *I, we y nation*
(θ_{diff}, prueba de proporción de Bayes mediante prueba exacta frente a fuerza bruta)

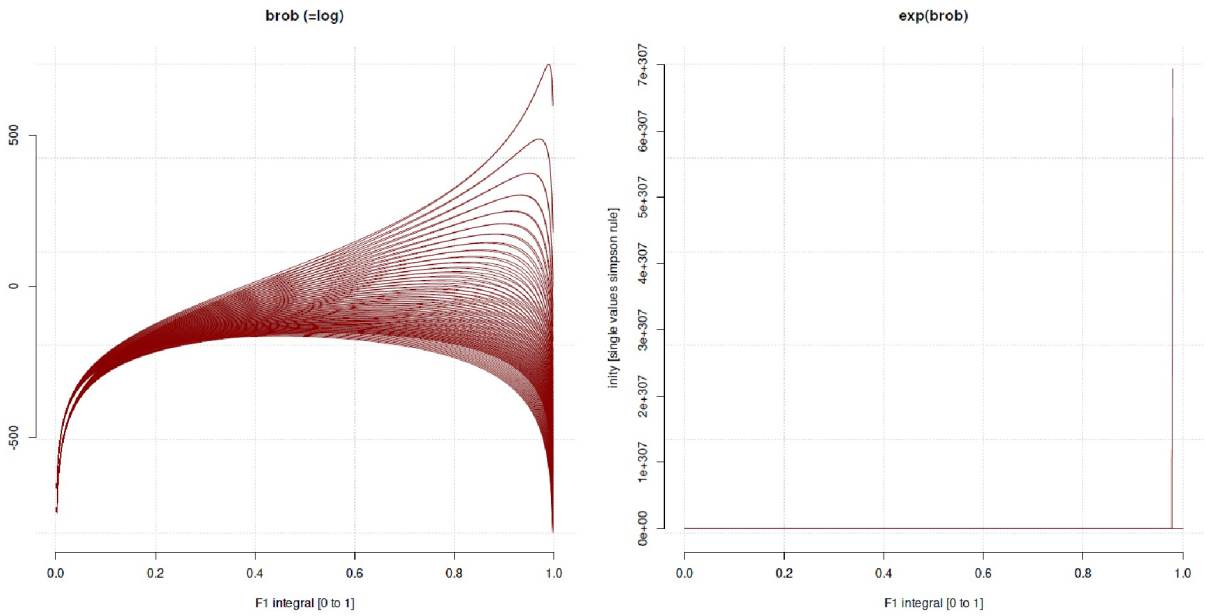


Figura 6.148. *I, we y nación* (θ_{diff}, *prueba de proporción de Bayes mediante prueba exacta, distribución hipergeométrica de integrales en 2D)*

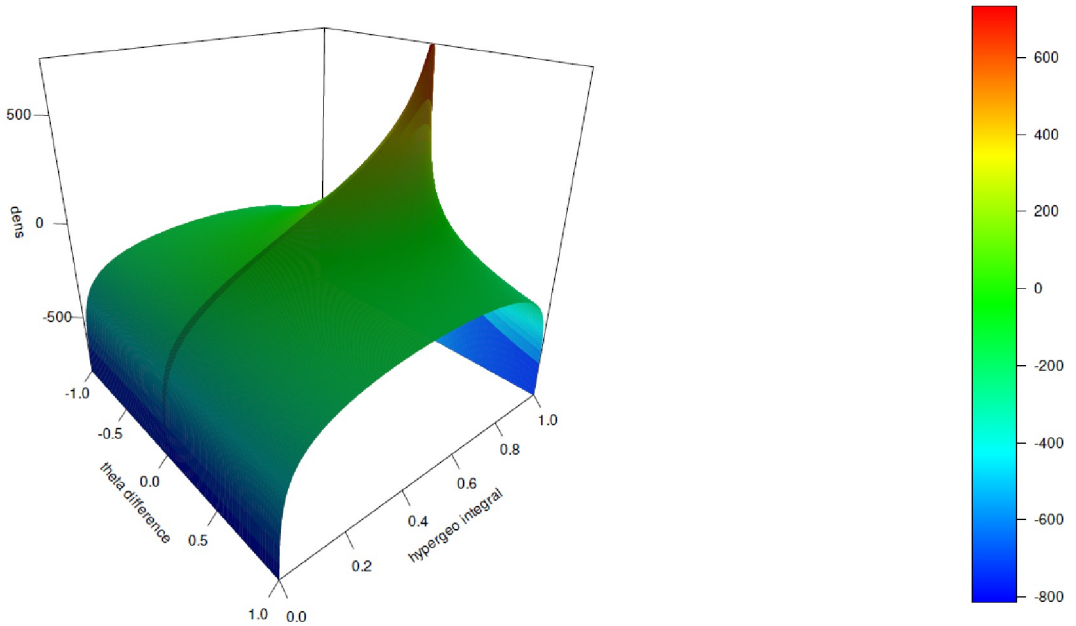


Figura 6.149. *I, we y nation* (θ_{diff}, *prueba de proporción de Bayes mediante prueba exacta, integrales de distribución hipergeométrica en 3D)*

Se pueden investigar las hipótesis individuales con subrutinas, por ejemplo si la diferencia de las proporciones es menor que un valor crítico θ_{crit}. Comparamos la fuerza bruta con la exacta mediante `cdf.theta.diff()`:

```
> # check values
> thetaC <- 0.05
> # brute force
> # p(difference < thetaC) = MCMC based
> mean(bf.diff < thetaC)
[1] 0.6176254
> # p(difference > thetaC) = MCMC based
> mean(bf.diff > thetaC)
[1] 0.382559
> # exact solution
> cdf.theta.diff(theta=thetaC, a1, b1, a2, b2)
[1] 0.6174502
```

La diferencia es prácticamente inexistente y así debe ser. Por último, veamos la hipótesis en contraste con las Posteriores completas de las diferencias de las propias proporciones – con `plot.bayes.prop.test.Xct()`. La función R toma un `data.frame()` de forma que la primera columna contiene la secuencia y la segunda los valores posteriores asociados. A esto se añaden los valores beta de las proporciones a_1 o b_1 y a_2 o b_2. Además de la salida gráfica de la diferencia de las proporciones, la función calcula las probabilidades de la diferencia, RR y OR en relación con un vector de valores críticos de comparación thetaCs, que recibe directamente los valores críticos que deben comprobarse para la diferencia, RR y OR. A continuación examinamos nuestros datos para una diferencia de 0,05, un RR de 0,85 y una OR de 1,15. Los valores que hemos elegido arbitrariamente y deberían justificarse en términos de contenido para ejemplos de investigación más serios.

```
plot.bayes.prop.test.Xct(res.Xct=Xprop.res[,c("sekk","pdf.diff")],
       a1=a1, b1=b1, a2=a2, b2=b2,
       thetaCs=c(0.05,0.85,1.15),
       loga=FALSE, drawmcmc=TRUE)
```

La figura 6.150 muestra el resultado gráfico; después sigue el resultado numérico (cada uno incluye la comparación con fuerza bruta):

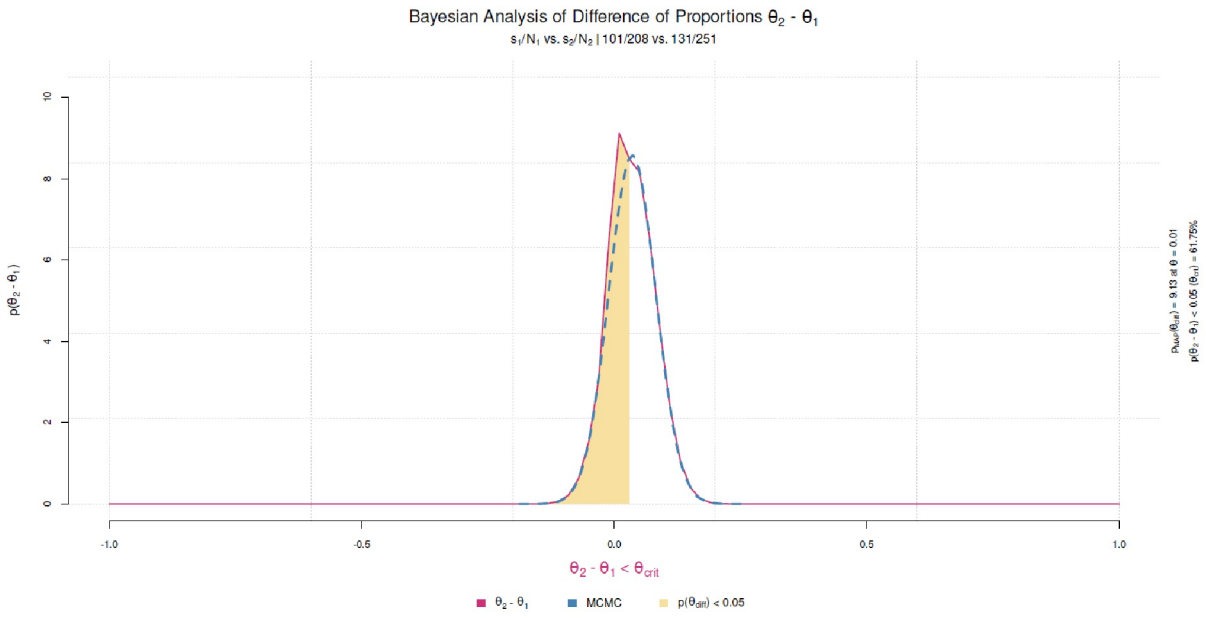


Figura 6.150. *I, we y nation (diff, prueba de proporción de Bayes mediante prueba exacta frente a fuerza bruta, marcación de la variable crítica de comparación)*

```
################################## R-Output
# Exact Bayesian Test of Proportions
### MAP Difference in Proportions:
Pr( [p2-p1] == max) Closest match
Exact: MAP = 9.1259 at theta difference = 0.0101
Brute Force: MAP = 8.5764 at theta difference = 0.036
### Difference in Proportions:
Exact:
[CDF] Pr( [p2-p1] < 0.05 ) = 0.62
[CDF] Inverse Pr( [p2-p1] > 0.05 ) = 0.38
[Ratio] ([p2-p1] < 0.05 ) / Pr( [p2-p1] > 0.05) = 1.62
Brute Force:
Pr( [p2-p1] < 0.05 ) = 0.62
Pr( [p2-p1] > 0.05 ) = 0.38
### Ratio of Proportions:
[CDF] Pr( [p2/p1] < 0.85 ) = 0.01
[CDF] Inverse Pr( [p2/p1] > 0.85 ) = 0.99
Ratio ([p2/p1] < 0.85 ) / Inverse Pr( [p2/p1] > 0.85) = 0.01
### Odds Ratio of Proportions:
[CDF] Pr( [p2/(1-p2)]/[p1/(1-p1)] < 1.15 ) = 0.49
[CDF] Inverse Pr( [p2/(1-p2)]/[p1/(1-p1)] > 1.15 ) = 0.51
Ratio ([p2/(1-p2)]/[p1/(1-p1)] < 1.15 /
Inverse Pr( [p2/(1-p2)]/[p1/(1-p1)] > 1.15) = 0.96
### Closest match of p(p2-p1 > crit) with crit = 0.05
       53
sekk 0.05050505
pdf.diff 8.16349807
### Closest match of p(p2-p1 < crit) with crit = 0.05
       52
sekk 0.03030303
pdf.diff 8.48830250
```

6.15.3.7 Factor de Bayes

Hemos ya informado de informes muy críticos sobre los factores de Bayes. Sin embargo, en nuestra opinión, las críticas se refieren a la forma en que se utilizan los factores de Bayes, al igual que ocurre con la estadística clásica y al problema de asignar *umbrales de significación globales* para los factores de Bayes comparables a los valores *p*, lo que traslada la discusión sobre la significación al ámbito bayesiano. El procedimiento en sí proporciona información como cualquier otro análisis y ésta puede ser o no útil y adecuada. En el caso de la prueba de proporción se dispone de dos funciones R del paquete R `BayesFactor` – `proportionBF()` y `contingencyTable()`. La primera prueba proporciones y la segunda tablas de contingencia. Para la prueba de proporciones se extraen los aciertos s y el número total de ensayos *N*. Como código R se obtiene (`ptII_quan_Bayes_case_presidential-debates.r`):

```
Si <- pres.2x2["I","Bush"]
Ni <- sum(pres.2x2[,"Bush"])
Sii <- pres.2x2["I","Kerry"]
Nii <- sum(pres.2x2[,"Kerry"])
```

Dejamos que la prueba de proporciones bayesiana pruebe contra $p = 0.5$, es decir, contra la hipótesis nula de que las proporciones están igualmente distribuidas.

```
bf.prop.pres <- proportionBF(y=c(Si,Sii), N=c(Ni,Nii), p=0.5)
# H1
bf.prop.pres
# H0
1/bf.prop.pres
```

Como era de esperar, esto refuerza los análisis anteriores. La hipótesis nula se ve claramente favorecida con un factor de Bayes de $BF_{H0} = 8.388$ frente a la hipótesis alternativa de una diferencia existente con $BF_{H1} = 0.119$. Es importante recordar que no se trata de una probabilidad posterior de la diferencia $\theta_1 - \theta_2$, sino la actualización de las expectativas previas tras la selección de los datos, de modo que la tesis "no hay diferencia" a la vista de los datos empíricos es más de 8 veces más probable. Las cadenas MCMC se podría ahora comprobar gráficamente aplicando la función de R `posterior()` directamente al objeto que contiene el factor de Bayes (el gráfico no se imprime).

```
nsims <- 1e5
pres.chains <- posterior(bf.pres, iterations=nsims)
plot(bf.prop.mcmc, col="darkred", bty="n")
```

No se detectan anomalías. El MAP de las Posteriores sería interesante. Este es casi exactamente MAP ≈ 0.5.

```
> # MAP
> dens.mcmc <- density(bf.prop.mcmc[,"p"])
> max.dens <- max(dens.mcmc$y)
> MAP <- dens.mcmc$x[dens.mcmc$y == max.dens]
> c("MAP"=MAP, "density"=max.dens)
MAP       density
0.5056806 17.5257726
```

La comparación obligatoria con las pruebas clásicas

```
prop.test(x=c(Si,Sii), n=c(Ni,Nii))
prop.test(x=c(Si,Sii), n=c(Ni,Nii), p=c(0.5,0.5))
```

muestra que clásicamente hay poco que aprender aparte de no rechazar la hipótesis nula. La primera prueba se realiza de forma no específica y la segunda bajo la especificación de probabilidades de la hipótesis nula de $p = 0.5$ (prueba bilateral). Las proporciones muestrales de $propB = 0.486$ para Bush y $propK = 0.522$ para Kerry son fácilmente compatibles con la hipótesis nula de $propH0 = 0.5$ y un valor p de $p = 0.496$.

La prueba de contingencia con `contingencyTableBF()` requiere dos datos adicionales sobre el plan de muestreo y las sumas marginales, que se explica en el manual de esta función de R. Nos decidimos por `sampleType`="indepMulti" y `fixedMargin`="cols". Es decir, el plan de muestreo sigue el supuesto de que los totales de los márgenes de fila o columna son fijos y cada fila o columna está distribuida multinomialmente. Bajo la validez de la hipótesis nula, se espera que las filas o columnas tengan las mismas probabilidades multinomiales. Como sumas marginales fijas, seleccionamos las columnas, es decir, Bush o Kerry, y no los términos "I" o "we/nation".

```
bf.pres <- contingencyTableBF(pres.2x2, sampleType="indepMulti",
fixedMargin="cols")
# H1
bf.pres
# H0
1/bf.pres
chisq.test(pres.2x2)
```

Se prefiere la hipótesis nula con $BF_{H0} = 6.341$ a la hipótesis alternativa con $BF_{H1} = 0.158$. Esto refleja la prueba de proporción anterior con `proportionBF()`, aunque aquí parece algo menos claro, pero no se debe sobreinterpretarlo. Al igual que con las soluciones del muestreador de Gibbs anteriores, se pueden generar extracciones de la Posterior, formar la frecuencia relativa para cada columna en relación con la suma

de columnas y a partir de esto formar la diferencia θ_{diff}, que a su vez se puede presnetar como gráfico (tampoco impresa).

```
nsims <- 1e5
bf.cont.mcmc <- posterior(bf.pres, iterations=nsims)
colnames(bf.cont.mcmc)
sametermsgivenBush <- bf.cont.mcmc[,"pi[1,1]"] /
       bf.cont.mcmc[,"pi[*,1]"]
sametermsgivenKerry <- bf.cont.mcmc[,"pi[1,2]"] /
       bf.cont.mcmc[,"pi[*,2]"]
theta.diff <- sametermsgivenBush - sametermsgivenKerry
mean(theta.diff)
# plot
par(oma=c(2,1,1,1), "cex.axis"=1, bty="l", mfrow=c(1,2))
plot(mcmc(theta.diff), main="", col="darkred", bty="n")
mtext(expression(paste("Bush vs. Kerry (",theta[1],"-",theta[2], ")",sep="")),
       outer=TRUE, line=-2, cex=1.5, side=3)
```

Las funciones gráficas del paquete BEST de R son más informativas (véase la Fig. 6.151). Aquí podemos añadir ROPE, un valor de comparación y un HDI.

```
plotPost(theta.diff, credMass=0.87, ROPE=c(-0.1,0.1),
       xlab="theta difference", showMode=TRUE,
       col="skyblue", border="white", compVal=0.25)
```

Por último, aparece de nuevo el valor MAP, esta vez el de la distribución posterior θ_{diff}.

```
> # MAP
> dens.mcmc <- density(theta.diff)
> max.dens <- max(dens.mcmc$y)
> MAP <- dens.mcmc$x[dens.mcmc$y == max.dens]
> c("MAP"=MAP, "density"=max.dens)
MAP        density
0.03763686 8.51350026
```

Esto tiene un valor de $MAP = -0.03$ con una desviación estándar posterior de $s_{post} = 0.047$

```
> sd(theta.diff)
[1] 0.04651868
```

y, por tanto, es menor que la dispersión posterior. Nuestras excursiones anteriores en el ámbito de la fuerza bruta también han reportado exactamente este valor de alrededor de 0.036 en el contexto de fluctuaciones aleatorias.

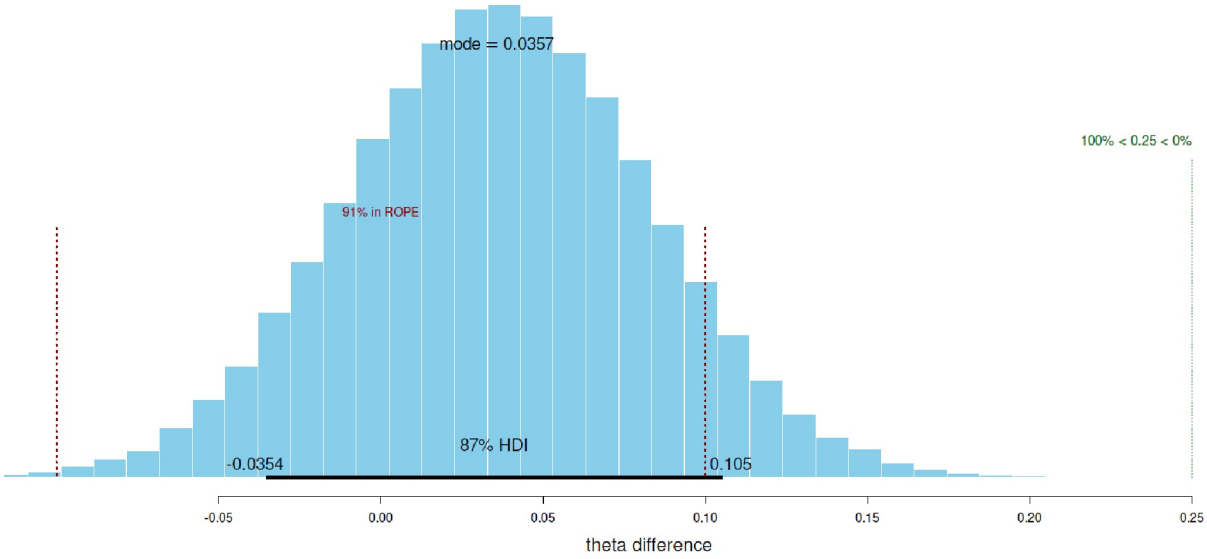


Figura 6.151. *I, we y nation* (*prueba de proporción de Bayes mediante factores de Bayes)*

6.15.3.8 Prueba t bayesiana según Bretthorst (1993)

En una interesante aplicación de la prueba *t* bayesiana en la implementación exacta según Bretthorst (1993), Studer (1996b, 1998) convierte las tasas de éxito en medias y desviaciones estándar que luego analiza utilizando la prueba *t* bayesiana para explorar las tasas de éxito en el tratamiento hospitalario de la adicción. Aquí aplicamos el procedimiento a los datos de los duelos orales. La función de R `SucRatesIntBounds()` convierte las tasas de éxito de cada caso en valores medios y desviaciones estándar y los prepara para el análisis con `DiffinMeans()`. Se basa en el trabajo de Studer (1996b, 1998). Dado que los cálculos intermedios se hacen muy grandes utilizamos el paquete `Brobdignang` de R, que nos permite trabajar sólo con logaritmos. La llamada correspondiente en `DiffinMeans()` es `BROB=TRUE`. El resto del código R se encuentra en (`ptII_quan_Bayes_case_presidential-debates.r`). Primero una recapitulación de los datos originales con sumas marginales:

```
> addmargins(pres.2x2)
Bush Kerry Sum
I 101 131 232
we/nation 107 120 227
Sum 208 251 459
```

Ahora preparamos el conjunto de datos y elegimos los límites superior e inferior de la Prior para la media y la desviación estándar. Para las tasas de éxito y el número total de casos utilizamos los mismos `Si`, `Sii`, `Ni` y `Nii` del ejemplo anterior de factores de Bayes.

```
res.SIB.pres <- SucRatesIntBounds(Si=Si, Ni=Ni, Sii=Sii, Nii=Nii, R-Code
smin=0,snames=c("I","we/nation"))
res.SIB.pres.upd <- res.SIB.pres
# "priors"
# less narrow, but not completely wide
```

```
# informed prior for means
# less informed for variances - more wide
res.SIB.pres.upd["L"] <- 0.4
res.SIB.pres.upd["H"] <- 0.6
res.SIB.pres.upd["sL"] <- 0.01
res.SIB.pres.upd["sH"] <- 0.1
```

A continuación figuran los cálculos reales

```
DiM.res.pres <- DiffinMeans(inval=res.SIB.pres.upd, out=FALSE, BROB=TRUE)
UMSprint(results=DiM.res.pres)
UMSplot(inval=res.SIB.pres.upd,pdfout=FALSE)
```

con la salida tanto numérica como gráficamente (véase la Fig. 6.152). Con `UMSplot()` se pueden trazar las distribuciones una al lado de la otra como se describe en Studer (1996b). Debe tenerse en cuenta que la figura representa las dos distribuciones investigadas y *no* la Posterior resultante de las hipótesis individuales. Esto se podría hacer utilizando las funciones de R `DiM.pg()`, `DiM.plot.calc.pg()` y `DiM.plot.pg()`. Encontrará información al respecto en el script de R `DiM_Bretthorst_PG_calls.r`.

```
######################################################################## R-Output
###
### ON THE DIFFERENCE IN MEANS
### G.L. Bretthorst (1993)
###
### original Mathematica code by U.M. Studer (90s, Switzerland)
Note:
If any probability is printed as '1' (= one) or '0' (= zero),
it means that the probability is practically that value by
giving respect to limited computer precision.
------------------------------- Data (Input) ------------------------
N_1 = 208 : Mean_1 ± SD_1 = 0.485714 ± 0.034407
N_2 = 251 : Mean_2 ± SD_2 = 0.521739 ± 0.031343
N_total = N_1 + N_2 = 459 : Mean_comb ± SD_comb = 0.5054 ± 0.0373
Bounds on the Mean (s_min = 0): Mean_L = 0.4, Mean_H = 0.6
Bounds on the Standard Deviation: SD_L = 0.01, SD_H = 0.1
Mean_L - Mean_comb < 0 = TRUE (-> '+'-sign between Gamma-fcts o.k.)
------------------------------- Results ----------------------------
...based on BROBs - try to convert back to non-BROB numbers...
p(mv | D_1, D_2, I) = const. +exp(1272.24)
p(mbarv | D_1, D_2, I) = const. +exp(1329.21)
p(mvbar | D_1, D_2, I) = const. +exp(1272.57)
p(mbarvbar | D_1, D_2, I) = const. +exp(1327.57)
where const. = 3.56792e-184 / p(D_1,D_2|I)
= +exp(-1329.39)
--------------- Model ---------------------------------- Probability --
mv: Same Mean, Same Variance: 1.50984e-25
mbarv: Different Mean, Same Variance: 0.836954
mvbar: Same Mean, Different Variance: 2.10697e-25
mbarvbar: Different Mean, Different Variance: 0.163046
---------------------------- Odds Ratios --------------------------
The probability the means are the same is: 3.6168e-25
The probability the means are different is: 1
The odds ratio is 2.76487e+24 to 1 in favor of different means.
The probability the variances are the same is: 0.836954
The probability the variances are different is: 0.163046
The odds ratio is 5.13322 to 1 in favor of the same variances
The probability the data sets are the same is: 1.50984e-25
The probability the data sets are different is: 1
The odds ratio is 6.62323e+24 to 1 in favor of different means
```

```
and/or variances.
------------------------------ End --------------------------------
#####################################################################
```

Si se compara la solución con, por ejemplo, la prueba clásica de χ^2 o la prueba de proporciones, que dan lugar cada una a un valor *p* de $p = 0.4957$ ($\chi^2 = 0.464$; df = 1), las afirmaciones sobre la relación entre media y varianza son aquí claramente más diferenciadas. Desde este punto de vista, los conjuntos de datos difieren muy bien, pero sólo en el valor medio y no en las varianzas. Mientras que la prueba de proporción examina la independencia multinomial de filas o columnas, la prueba *t* exacta separa el problema en los componentes básicos de la prueba *t* de Behrens-Fisher.

Para más análisis se pueden desplazar los límites a priori cambiando las variables `L` y `H` o `sL` y `sH`

```
res.SIB.pres.upd["L"] <- 0.4
res.SIB.pres.upd["H"] <- 0.6
res.SIB.pres.upd["sL"] <- 0.01
res.SIB.pres.upd["sH"] <- 0.1
```

para, a continuación, repetir el procedimiento:

```
DiM.res.pres <- DiffinMeans(inval=res.SIB.pres.upd, out=FALSE, BROB=TRUE)
UMSprint(results=DiM.res.pres)
```

que en la aplicación proporciona declaraciones actualizadas sobre la diferencia de medias y varianzas.

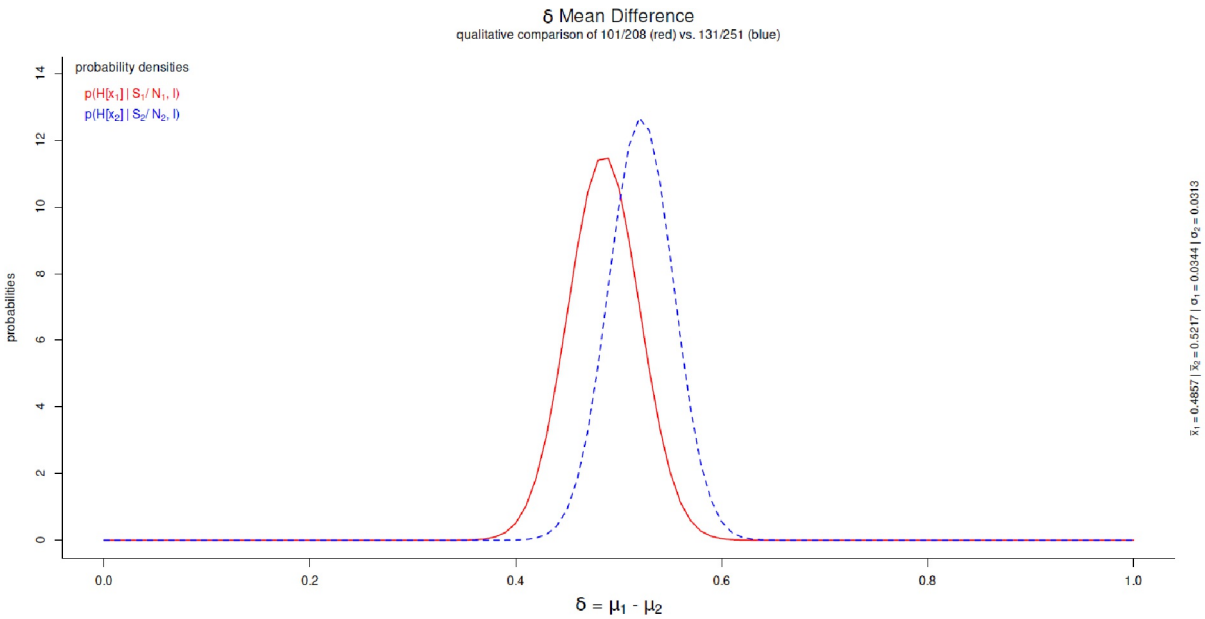


Figura 6.152. *I, we y nation (prueba de proporción de Bayes mediante la prueba t bayesiana exacta según Bretthorst, 1993 tal como utilizado en Studer, 1996b)*

6.15.3.9 Resumen

Ninguno de los análisis binomiales o de proporciones, clásicos o bayesianos, ni los resultados gráficos aportan pruebas de que hubiera diferencia alguna en el uso de los términos "I" y "we/nation" difirió sustancialmente entre el presidente Bush y el aspirante Kerry durante los tres duelos de discursos de la

campaña presidencial. Sin embargo, el análisis según el enfoque de Bretthorst (1993) tras la aplicación de Studer (1998) llega a conclusiones diferentes.

En este caso, los valores medios difieren significativamente entre sí (tras convertir los porcentajes de éxito en valores medios y desviaciones estándar), pero no las varianzas. La razón del resultado supuestamente diferente es que de esta forma se examinan hipótesis completamente distintas. No se trata de la comparación de dos proporciones, sino del análisis separado y combinado de las medias y las desviaciones estándar de estas proporciones. Las hipótesis resultantes arrojan un panorama más diferenciado. Desde esta perspectiva, Kerry se dirigiría más claramente al "yo" que Bush - comparando con el "nosotros/ nación", mientras que la variabilidad en el uso de los términos es comparable, si no idéntica, en ambos. Una prueba pura de la diferencia en los porcentajes de éxito no puede hacer tales afirmaciones, por lo que los análisis no son directamente comparables. Todo esto no significa que los discursos fueran idénticos ni que fueran estructuralmente equivalentes. Tampoco dice nada sobre si un análisis detallado de la terminología utilizada no podría conducir a resultados bastante diferentes.

Recordatorio 6.6: La pregunta determina la respuesta

Se puede observar que la pregunta de investigación y, en consecuencia, la elección del procedimiento analítico, condicionan inevitablemente la naturaleza de la respuesta empírica y, por tanto, el espacio de posibilidades de los resultados. Lo que no preguntamos y donde no permitimos una respuesta, no lo obtenemos y no debemos sorprendernos después. Lo que no preguntamos, pero permitimos una respuesta, ahí podemos obtener respuestas.

Los términos se evaluaron a ciegas y sin contexto. Se trata de un análisis interesante pero poco elaborado que podría mejorarse considerablemente en el aspecto cualitativo. Desde un punto de vista cuantitativo, se podría formular un modelo lineal de Poisson de diversas características de los discursos, subdividido según los tres momentos y para términos específicamente seleccionados, con el fin de determinar las posibles diferencias entre los dos oradores. Esta sería una tarea para los lectores interesados.

Si desea hacerlo de forma exhaustiva, puede ejecutar un modelo de este tipo sobre todos los discursos documentados hasta la fecha de las campañas electorales presidenciales estadounidenses, de modo que surja entonces el análisis de un modelo lineal jerárquico, que también puede examinar la evolución a lo largo del tiempo.

Sin embargo, lo que es mucho más importante es que investigamos lo mismo con una amplia variedad de métodos de análisis de datos, y los métodos que trabajan sobre la misma hipótesis también llegan a un resultado muy comparable – independientemente de si se trata de fuerza bruta, un muestreador de Gibbs o soluciones exactas. Sin embargo, si a continuación examinamos una cuestión diferente, no debería sorprendernos si también obtenemos una respuesta diferente. En ese caso, depende de la interpretación del contenido como se puede integrar esta información.

6.16 Hacer accesible el conocimiento experto a las distribuciones a priori

Por último, dedicamos un capítulo al libro de O'Hagan, Buck, Daneshkhah, Eiser, Garthwaite, Jenkinson, Oakley y Rakow (2006). Consideramos el libro tan importante que le dedicamos un breve capítulo. Hasta donde sabemos, los autores han escrito el único libro sobre estadística de Bayes que trata explícitamente de la construcción de distribuciones a priori, y no como una larga serie de argumentos matemáticos que no

tienen nada que ver con el contenido e ignorando lo esencial. Las distribuciones a priori no tratan precisamente de propiedades matemáticas preferidas que uno considera separadas del contenido, como hacen varios defensores de la estadística de Bayes. Más bien, el objetivo es demostrar que una distribución a priori representa de forma adecuada y precisa el estado actual del conocimiento. Si luego se añaden propiedades matemáticas favorables, por supuesto que es de agradecer, pero en este contexto el contenido es más importante que las matemáticas. El contenido siempre tiene prioridad sobre las matemáticas, y esto debe considerarse así sin excepción, ya que las matemáticas son una herramienta y no un fin en sí mismo.

La esencia de la distribución a priori se divide entre dos grupos de cuestiones:

- ¿Existen estudios empíricos pasados cuyos resultados puedan utilizarse directamente como Priores? ¿Se pueden poner en forma numérica los resultados de estos estudios? ¿Es posible derivar afirmaciones concretas sobre el aspecto, los rangos de valores esperados y, posiblemente, la regularización de la Prior?
- Si no existen estudios empíricos sobre el tema: ¿Cómo se puede aprovechar el conocimiento experto para llegar a una Prior justificada y razonable?

Los autores dedican especial atención a la segunda cuestión: cómo obtener conocimientos y distribuciones a priori de los expertos de forma estructurada. En sentido estricto, no se trata de un enfoque independiente del análisis de datos, sino de una descripción concreta de los pasos de este proceso. Los autores describen este proceso con gran relevancia práctica, según distintos puntos de vista, y ofrecen ejemplos y listas de comprobación. Prestan especial atención a cómo introducir a los profanos y a quienes no familiarizados con las matemáticas en los conceptos de la estadística bayesiana y la noción subyacente de probabilidad, y cómo utilizarlos como base para pensar y sacar conclusiones para la extracción de datos.

Además, abordan las trampas psicológicas de la percepción y los sesgos perceptivos típicos que llevan a integrar la información de forma incorrecta y, por tanto, a llegar a conclusiones falsas en el contexto de una información incompleta. Abarcan tanto el caso discreto como el paramétrico y abordan diferentes escenarios y grupales en la utilización del conocimiento experto. A continuación se evalúan las Priores producidas y se estudian diversos casos de los campos de la medicina, la industria nuclear, la meteorología, la economía, la agricultura y la veterinaria. Los autores defienden la necesidad de un software específico para poner el conocimiento experto a disposición común después de haberlo hecho consciente y desprendido de los diversos contextos temáticos en una forma matemática, que a su vez puede ejercer su influencia natural en el marco de un análisis bayesiano de datos.

Sin embargo, el libro necesita un complemento. Por ejemplo, se describe la forma en que se hace consciente el conocimiento experto anclado en el contexto, pero se echa de menos una estructura aún más científica de este proceso antes de que pase al ámbito matemático. Por ejemplo, se podría utilizar el análisis secuencial de la hermenéutica objetiva tras recopilar información previa, para ponerla en forma estructurada, es decir, poner la información en una forma estructurada antes de pasar a una distribución matemática. Un procedimiento de este tipo tendría la ventaja de que la información no se obtiene, se hace accesible y se procesa de forma arbitraria, sino de manera estrictamente científica. También se podría y se debería conservar el paso de la falsificación, con el fin de examinar críticamente la información extraída y condensada en una estructura de casos para determinar su validez como representante del conocimiento previo. Una implementación matemática podría entonces ser quizás aún más fácil y específica.

En resumen, este libro representa una ampliación significativa del arsenal estadístico bayesinano común. Demuestra de forma impresionante que es cuestión de combinar información cualitativa y consideraciones matemáticas si se quiere utilizar la información a priori como distribuciones a priori.

6.17 Conclusión: La estadística bayesiana

El capítulo anterior contiene suficientes puntos de partida para formarse una imagen propia de la estadística bayesiana. Estos puntos no se repetirán ahora. ¿Cuáles son los argumentos a favor de la estadística bayesiana? En nuestra opinión, Bayes es un instrumento generalmente aplicable, flexible e intuitivamente comprensible para tratar las probabilidades, que permite realizar todos los análisis de la estadística clásica. Permite incorporar tanto nuestro conocimiento de los contextos empíricos como tener debidamente en cuenta la incertidumbre asociada a la investigación respecto a los datos y los modelos. Cómo es la realización en casos concretos no siempre está claro, es decir, existen (casi) siempre diferentes soluciones posibles. Esta incertidumbre refleja la perpetua verdad relativa en el ámbito sensiblemente perceptible y corresponde al marco epistemológico que se defiende aquí en el libro.

La relatividad de la verdad no nos exime de la necesaria garantía de calidad. En este sentido, la siguiente estrategia es actualmente la más prometedora desde nuestro punto de vista. Incluye un resumen de los argumentos ya enumerados:

- Implementar un modelo bayesiano completo que conste de Prior, Likelihood y Posterior.
- Especialmente en el caso de un número reducido de casos y cuando se disponga de información contextual o de estudios empíricos sobre el objeto de investigación, se debería llevar a cabo una reconstrucción cualitativa metodológicamente controlada. El objetivo es desarrollar una estructura bien fundamentada de la que se pueda derivar una Prior. Sugerimos el análisis de secuencias (véase el capítulo 11.2) como método de elección.
- Modelización compleja en lugar de pruebas aisladas, es decir, no centrarse (sólo) en factores de Bayes o en modelos y variables individuales que sean superiores a otros. Éstos sólo permiten responder *sí/no* a preguntas de investigación limitadas y, naturalmente, ignoran las preguntas relevantes sobre el *cómo*. Por el contrario, los modelos complejos deben adaptarse de modo que los enfoques explicativos que compiten entre sí se integren como casos especiales en el modelo complejo. La dialéctica hegeliana como proceso dirigido a objetivos parece ser una buena base para esta actitud investigadora. Consideramos problemática la aplicación exclusiva de la lógica aristotélica del *"o bien"* o de figuras puras de *"si-entonces"*, ya que éstas también son sólo construcciones y la realidad no funciona necesariamente así, sobre todo cuando se trata de sistemas vivos.
- Uso habitual de métodos gráficos para evaluar la calidad de los modelos. Los coeficientes por sí solos no bastan. Los métodos gráficos deben utilizarse tanto para el AED como para la comprobación de modelos.
- Se necesitan comprobaciones posteriores predictivas y, si es posible, réplicas para garantizar que los ajustes del modelo no son arbitrarios.
- La replicación de estudios permite contrastar los modelos estimados con nuevos datos y ajustarlos si es necesario. Cuando el modelo no se ajusta, es cuando más hay que aprender. Esto requiere una reorientación fundamental de las revistas pertinentes, que dejen de lado la importancia y se centren en una evaluación exhaustiva de los informes de investigación y un tratamiento igualitario de los resultados menos "significativos" y de los "fracasos" frente a los "informes positivos". El criterio para una publicación, además de la seriedad, la transparencia, el cumplimiento de las normas científicas, etc. debería ser que se pueda aprender algo de un estudio y que se pueda avanzar en un área temática como resultado. Sin duda, esto no cambiará muy rápidamente, ya que la mayoría de las revistas están dominadas por la idea olímpica de la cuantificación en lugar de un enfoque reflexivo sobre los requisitos de la ciencia. Mientras las revistas se nieguen a publicar intentos de réplica, como en el estudio Bem (véanse los capítulos 4.4.2.2 y 6.8.1.6), o mientras exista un fuerte sesgo que favorezca los estudios con resultados positivos frente a los negativos, las revistas científicas no serán realmente una comparación realista de las actividades de investigación.
- Además, está la repetidamente mencionada falta de potencia de la mayoría de los estudios, un problema bien estudiado y conocido de la investigación en ciencias sociales desde hace décadas. Lo mismo ocurre con la preferencia por los estudios con resultados "significativos". Afortunadamente, está aumentando el número de publicaciones que abordan precisamente estos problemas.

Especialmente para los investigadores sin conocimientos matemáticos profundos, la estadística de Bayes se puede convertir rápidamente en algo muy exigente. Esta situación se ha mejorado significativamente por

la disponibilidad de muchos paquetes de R (`BEST, brms, rstanarm`, s.a.CRAN, 2019a) o el programa `JASP` (2018), que se basa en R. La realización de los cálculos – si se elige una distribución a priori adecuada – es igual de fácil de realizar que cuando se utiliza estadística clásica en R. En muchos casos, incluso se puede conservar la notación familiar de R. Además, ya están disponibles toda una serie de publicaciones de alto nivel que describen la implementación de la estadística bayesiana en `R` (2019d), `Stan` (2019b), `JAGS` (2019) o `BUGS` (2019) utilizando muchos ejemplos reproducibles que pueden adaptarse a las propias necesidades. Sin olvidar los innumerables blogposts, tutoriales, materiales didácticos y de cursos, mensajes en foros, etc. que hacen lo mismo. Aunque en principio toda contribución debería ser examinada en cuanto a seriedad y calidad, estas contribuciones informales están muy a menudo al mismo nivel que los artículos de revistas. Estas últimas no están en absoluto exentos de errores, sino todo lo contrario, como demuestra de forma impresionante el estudio de Bem (2011a) sobre la clarividencia psíquica. Desde luego, no se trata de un caso aislado, porque ningún estudio es perfecto. Se plantean dos escenarios:

- La publicación en una revista (de prestigio) ya no es suficiente para aceptar una investigación como seria y bien planteada per se.
- El rechazo de un artículo en una revista (reputada) tampoco es garantía de saber si un estudio o trabajo aporta valor científico o no.

Para nosotros, esto significa que todos los trabajos deben tratarse con cautela y que siempre hay que leer los artículos de forma crítica. A la inversa, sin embargo, no significa que no se pueda aprender nada de un artículo rechazado. Durante la elaboración de este libro, hemos leído algunos artículos rechazados que nos han parecido excelentes. A veces también se trata del espíritu de la época: por ejemplo, el físico E.T. Jaynes tuvo problemas durante mucho tiempo para publicar sus trabajos en revistas de física, y eso que a pesar de la calidad (ahora indiscutible) de su trabajo.

Si esto no es suficiente, le recomendamos que busque a alguien que conozca la estadística bayesiana y trabaje en estrecha colaboración con esta persona. Esta persona debería, además de matemática subyacente, tener una sólida experiencia y práctica en cómo el conocimiento experto implícito es explorado, razonado, reconstruido y transformado en Priores (O'Hagan, Buck, Daneshkhah, Eiser, Garthwaite, Jenkinson, Oakley & Rakow, 2006). Porque de eso trata Bayes. Como investigadores, es importante no perder de vista el empirismo – sin embargo, no es necesario llevarlo todo a cabo uno mismo hasta el último detalle. Nosotros mismos preferiríamos de inmediato trabajar con un bayesiano excelente en lugar de hacerlo solamente por nuestra cuenta y nuestra práctica (bastante sólida con la estadística clásica o bayesiana).

Capítulo 7
Discusión: Las Estadísticas

»Reports that say that something hasn't happened are always interesting to me, because as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns — the ones we don't know we don't know.«

Pressekonferenz, Februar 2002
Donald Rumsfeld, 1932–

7.1 Interpretación de las estadísticas

La interpretación de las estadísticas no debe confundir ni mezclar los distintos niveles del proceso científico. Si lo hace, las múltiples traducciones quedarán más o menos distorsionadas. El grado de distorsión puede ser tan indeterminado, en función de su localización, que no sea posible cuantificarlo ni aprehenderlo cualitativamente.¿Se puede hablar entonces de distorsión? La respuesta es algo así como la pregunta de si podemos elevarnos por encima del sistema en el que residimos. No obstante podemos y debemos intentarlo:

- Concepción científico-teórica del mundo
- Diseño de la investigación
- Muestra y datos
- Matemáticas puras
- Estadística
- Conclusiones y causalidad

7.1.1 Cosmovisión de la filosofía de la ciencia

Es el punto de partida de todo y en realidad – dado un criterio relativo de verdad – una cuestión subjetiva de fe, incluso para los científicos, aunque no les guste oírlo o admitirlo. Los axiomas o una comprensión teórica científica básica no pueden ser definitivamente demostrada. Esto lo sitúa en el nivel de una cuestión de fe, aunque parezca lógicamente justificable. Las premisas no pueden demostrarse definitivamente. Para quien la tierra no es redonda, probablemente no encontrará nada que demuestre lo contrario, porque no se pide. Se trata de premisas implícitas, como por ejemplo cómo se definen conceptos como objetividad desde

el punto de vista de investigadores subjetivos, etc., o qué criterios de calidad se consideran válidos. Muchas veces se describe algo según convenciones sin cuestionarlas (véase la crítica discusión del significado de la estadística clásica en muchos lugares del libro, por ejemplo, en los capítulos 4.3.9.1 y 4.3.8). Sólo una meticulosa investigación sería capaz de averiguar los supuestos implícitos que son principalmente eficaces, si es que los hay. Por tanto, la transparencia y la metacomunicación son, en nuestra opinión, importantes criterios de calidad científica, porque no podemos trabajar sin una visión del mundo, pero debe quedar claro para el discurso a cuál nos referimos coherentemente.

7.1.2 Diseño de la investigación

El diseño general de la investigación (por ejemplo, experimento de laboratorio, estudio de campo, etc.) determina si un estudio pretende generar o probar hipótesis o modelos. Esto debe derivarse directamente de la pregunta de investigación general. Las estadísticas ofrecen pistas, pero no determinan el diseño. El diseño es general, las estadísticas, al igual que todos los instrumentos de investigación y procedimientos de análisis de datos son auxiliares y subordinados. Un plan de investigación debe reflejar esto de tal forma que la naturaleza y la dirección de las hipótesis y el objetivo básico – por ejemplo, exploración, confirmación, replicación – sean siempre centrales. Si se confunde la jerarquía tomando decisiones basadas en la matemática y no en motivos de diseño, pueden surgir distorsiones e incoherencias en las conclusiones.

7.1.3 Muestra y datos

Los datos en sí no deciden lo que ocurre con ellos. Sin embargo, sí proporcionan las limitaciones qué posibilidades cuantitativas son factibles. El tamaño de la muestra y una selección inteligente de los parámetros del modelo permiten encontrar correlaciones. Si no se recogen el sexo y la edad, no se pueden encontrar correlaciones con estas variables. Si la muestra no representa las proporciones reales de sexo y edad en la población de referencia, los resultados sólo pueden aplicarse a poblaciones de forma limitada, hasta cierto punto, etc. El estudio de Wipfler (2017) del que se informa en el capítulo 5.5.6 muestra de forma impresionante las consecuencias si la selección de la muestra es desfavorable.

7.1.4 Matemáticas puras

Las matemáticas en sí no pueden evitar ni defenderse de lo que se le hace. Un valor *p* no es más responsable de ello que una probabilidad posterior, la interpretación de los valores resultantes y su clasificación. Los algoritmos per se están libres de contexto, pero nunca su aplicación. En consecuencia, existen diferencias entre los partidarios de la estadística frecuentista y la bayesiana, no son las matemáticas y los algoritmos los que se ponen en duda mutuamente, sino las interpretaciones y conclusiones de los resultados.

7.1.5 Estadística

La elección de los métodos estadísticos determina el tipo de conclusiones y el tipo de información que entra en las ecuaciones y es procesada por ellas. La elección en sí no es necesariamente inequívoca, de modo que sólo una variante o un modelo sea el correcto. Más bien la práctica estadística consiste en un cierto ensayo y error para encontrar un modelo adecuado. Es importante argumentar de forma coherente en la línea de la teoría estadística elegida. Esto puede dar lugar a problemas, por ejemplo, si no se puede responder en absoluto a preguntas importantes o sólo se responde de forma inadecuada con el enfoque elegido. La

significación estadística en cualquiera de sus formas se deriva de la elección de criterios, que no tienen qué ser de carácter estadístico, sino de contenido – a menos cuando se decida aplicar límites convencionales que no estén justificados en la materia – y que desaconsejamos.

7.1.6 Conclusiones y causalidad

Un análisis cuantitativo no puede hacer afirmaciones directas sobre la causalidad de las relaciones causa-efecto, sino que sólo proporciona indicios. La causalidad en sí es – si acaso – visible solamente a través de un diseño de investigación bien planteado, en el que seguramente el empirismo (diseño, análisis de datos, etc.) desempeña el papel central. A esto se añade la necesidad de replicación: la causalidad no sólo debe ser corroborada, sino también replicable. Idealmente, la causalidad surge a través del diseño experimental. Sin embargo, hay contextos que no se pueden investigar experimentalmente, como las supernovas en astrofísica. No obstante, se pueden establecer mecanismos causales de causa y efecto. Sin embargo, esto requiere un gran esfuerzo.

7.2 La elección del enfoque

Ahora se plantea la siguiente cuestión: ¿Cómo debemos trabajar estadísticamente en concreto? ¿Qué enfoque debemos seguir? ¿Ha demostrado ser claramente superior a los demás? Se han hecho muchos comentarios críticos sobre la estadística clásica y no podemos ocultar un claro afecto por la estadística bayesiana. Pero la estadística bayesiana no está en absoluto exenta de problemas, como demuestran las observaciones sobresubjetividad, la elección de la Prior o de los factores de Bayes y la tendencia humana básica a la (semi)-automatización. Como ocurre a menudo, depende de qué cuestión se vaya a investigar estadísticamente, de qué competencias se disponga para realizarla, y si se elige un procedimiento que se ajuste realmente al caso. Gigerenzer y Marewski (2015) advierten contra la adhesión al ideal del análisis (semi)automatizado de datos estadísticos. Hay poco que añadir a esto.

Como se ha mostrado en el capítulo 4.3.3.4, el enfoque Neyman-Pearson encaja muy bien en ámbitos como la la gestión de la calidad. La estadística bayesiana (véase el capítulo 6) es especialmente interesante cuando se trata de incluir conocimiento contextual y la probabilidad de las hipótesis está en primer plano en lugar de la de los datos. Del mismo modo, todo el proceso bayesiano de estimación y predicción de modelos es ahora extremadamente potente gracias a la tecnología informática disponible y a las posibilidades de visualizar procesos y resultados mediante gráficos – y se puede aplicar Bayes realmente en cualquier lugar y en cualquier momento.

Incluso Fisher (véase el capítulo 4.3.2) tiene su lugar, a saber, cuando sabemos poco y un área es desconocido para nosotros, no tenemos preparada una explicación alternativa y necesitamos una orientación inicial aproximada. Pero no nos detenemos en ese punto, por lo que el alcance de Fisher termina bastante rápido.

El único enfoque para el que no tenemos comprensión ni vemos espacio en la estadística es la prueba de significación de hipótesis nula, aún muy extendida (véase el capítulo 4.3.8), con la exageración excesiva del valor *p*, el umbral crítico de significación elegido únicamente por convención y la falta de integración de la teoría de Fisher y Neyman-Pearson. Por desgracia, esto sigue dominando los libros de texto, la enseñanza y determina las posibilidades de publicación en muchas revistas. Increíblemente, teniendo en cuenta que hace más de 20 años la APA comenzó a desaconsejarlo muy claramente e incluso lejos de la estadística de Bayes las críticas se remontan a la década de 1960. Por no mencionar que el difunto Fisher ya advirtió en los años 50 en contra de la estadística de Bayes.

Fisher ya advirtió en la década de 1950 del peligro de utilizar siempre los mismos umbrales de significación sin pensar. Sin embargo, si por la razón que sea, la estadística clásica debe ser el enfoque principal, las siguientes recomendaciones pueden ser útiles, que, por cierto, se aplican igualmente a la

estadística bayesiana. Aquí nos guiamos por el modelo de investigación cíclica (véase el capítulo 2):

7.3 La pregunta de investigación

La pregunta de investigación es el Alfa y el Omega según nuestra forma de entender la ciencia. Los métodos, los instrumentos de encuesta y los análisis son accesorios necesarios, pero no persiguen un fin en sí mismos más allá de eso. La conducción al empirismo – desde la pregunta de investigación -- y el retorno de los resultados a ella constituyen el núcleo del proceso científico. Dado que la investigación consiste en procesos ininterrumpidos de traducción entre teoría, empirismo, suposiciones e hipótesis, datos y referencias, cifras y textos, etc., la pregunta de investigación constituye el hilo "rojo" de conexión. Lo vemos independientemente de si estamos investigando la causalidad lógica, tenemos ante nosotros un laboratorio experimental, un estudio de campo exploratorio o la derivación de intervenciones prácticas en un entorno terapéutico. Básicamente, siempre se trata de relaciones causa-efecto o, al menos, de relaciones correlativas, si es que las relaciones causa-efecto aún no están del todo claras. En el proceso, utilizamos constantemente modelos e ideas que no deberíamos tomar demasiado en serio porque, de todos modos, no son ciertos y cambiarán. Sin embargo, la pregunta de investigación que nos acompaña constantemente nos permite preguntarnos en cada paso "¿En qué medida este paso nos ayuda a responder a nuestra pregunta de investigación?". Esto no significa en absoluto ser poco creativo y, en consecuencia, evitar nuevas formas de hacer las cosas. Todo esto lo vemos independientemente de si trabajamos o no estadísticamente y, en caso afirmativo, según cuál de las variantes presentadas.

7.4 Documentación

Las fijaciones por escrito de hipótesis, expectativas, interfaces entre empirismo y teoría o viceversa, etc. permiten reproducir exactamente los estudios. Además, favorecen un trabajo serio, ya que la documentación impone un cierto orden. Y la documentación libera de la confusión para poder concentrarse mejor en los datos empíricos. Si hay suposiciones claras sobre, por ejemplo, correlaciones que se van a "probar", todas ellas deben establecerse por escrito de antemano. Lo mismo se aplica a los tamaños de las muestras y a lo contrario de lo que se espera, de modo que no se busque únicamente la confirmación de las propias suposiciones. Cuanto más haya por escrito mejor. Esto evita que algunos hagan "accidentalmente" el corte en la recogida de datos para que los resultados parezcan buenos – porque son significativos – o que se apeguen demasiado a sus propias expectativas.

7.5 Salida de datos completos y AED

Esto implica la salida de todas las estadísticas descriptivas, así como las intercorrelaciones de las variables investigadas, además del examen gráfico de los datos brutos en términos del AED de Tukey (véase el capítulo 5). Todo lo que muestra el EDA debería ser obvio antes de la validación estadística posterior. Si existen discrepancias claras entre el AED o las expectativas generadas por él y los análisis estadísticos, éstas deben investigarse de tal forma que las diferencias resulten comprensibles. Los gráficos útiles son los de dispersión, los de densidad en combinación con histogramas, diagramas de interacción, diagramas con errores estándar implementados, así como diagramas de caja – y siempre para el total respectivo y los subgrupos de interés.

Para muchos grupos, los gridplots (paquetes de R `grid, gridExtra, lattice` y `ggplot2`) son otra ayuda indispensable.

7.6 Renuncia a los rituales y apertura a la flexibilidad

Abandonar las barreras rígidas de significación que no se justifican en términos de contenido ni son justificables de ninguna otra manera conduce a una mayor reflexión. Si se comunican valores p, hay que relacionarlos con su significación práctica, su potencia, el tamaño de la muestra y su importancia en la realidad. Pensamos que establecer límites por convención es mala ciencia. Entonces todos sabemos a qué nos referimos y nos lo aseguramos mutuamente, pero no tenemos ni idea de lo que esto significa para la realidad. La discusión sobre HLMs/MLMs (Bates, 2006) muestra que el cálculo de estos valores tan importantes no sólo no es trivial, sino que en algunos casos apenas es posible o es propenso a errores, por lo que su uso está prohibido. Sacar conclusiones basándose en parámetros propensos a errores no parece muy científico. A este respecto, el trabajo de Andrew Gelman y sus colegas nos parece ejemplar. En general, se pueden derivar más directrices para practicar la estadística con sensatez:

- Reflexión sobre cómo formular las hipótesis (estadísticas) H_0 y H_1 – o lo que sea pertinente – y qué consecuencias concretas se derivan de la(s) hipótesis(s).
- Estimación de los parámetros de los modelos en lugar de pruebas estrictas de los mismos.
- Predecir modelos sobre nuevos datos o al menos sobre simulaciones
- Formulación de modelos complejos que integren casos especiales en lugar de limitarse a probar modelos entre sí
- HLMs/ MLMs en lugar de simples regresiones y anovas
- Abandone completamente las declaraciones de significación, si es posible abandone los *p-valores*, en su lugar considere los tamaños del efecto y discútalos siempre en su escala original
- Atención a los errores de tipo S y de tipo M (dirección y tamaño de los efectos) y a los errores de tipo III y de tipo IV (teoría y modelo de trabajo, así como la conexión entre los supuestos del modelo y los resultados)
- Inspección gráfica y exploración de modelos, residuos e interacciones en lugar de fijarse sólo en los coeficientes.
- Replicación, ya que siempre puede haber efectos de muestreo (considere el equilibrio entre aprender de investigaciones anteriores y seguir replicando lo suficiente como para que sea realmente una replicación).

Como ya se ha indicado (por ejemplo, véase el capítulo 6.8.2), varios autores sugieren utilizar criterios cualitativamente diferentes para evaluar los modelos. En el caso de los modelos de ecuaciones estructurales, éstos incluyen los índices AIC, BIC, RMSEA o GOF (Brandstätter, 1999) en lugar de los *p-valores*. Se puede adoptar un enfoque equivalente para otros modelos. Pero aquí también es importante considerar las ventajas y desventajas (por ejemplo, dependencias de la muestra, independencia/dependencia del número de parámetros del modelo, ...) de las variables respectivas. *Diferente* no siempre es *mejor* y el mejor criterio no existe.

7.7 Perspectivas múltiples

Utilizar diferentes técnicas de análisis de datos sobre el mismo problema ayuda a comprender la gama de efectos. Esto permite observar el comportamiento de los datos y los parámetros estimados en diferentes

condiciones. Esto es mejor que suponer que existe un único modelo mejor. Nos permite estudiar cómo los datos se comportan de forma diferente en función del modelo y qué modelos pueden o no explicar algo y por qué. A partir de ahí, se puede aprender y seguir desarrollando los modelos. Esto comienza con el uso de HLMs/MLMs (paquetes R lme4, nlme), incluso si "realmente" sólo se trata de un anova con repetición de medidas. Gelman y Hill (2007) recomiendan como regla general utilizar HLMs/MLMs si es posible desde el punto de vista de los datos. Los HLM/MLM son mucho más robustos que un anova simple y permiten la introducción de estratos, tatar factores de influencia significativos, limpiar los efectos y elaborarlos de forma más clara. Asimismo, en el caso de HLMs/ MLMs, por ejemplo, merece la pena recurrir adicionalmente a variantes robustas para investigar la constancia de los efectos encontrados (paquete R robustlmm). No obstante, se recomienda precaución para garantizar que el cambio de procedimiento sigue examinando las mismas hipótesis. El cambio a un procedimiento basado en rangos sólo examina entonces las diferencias en el nivel de rango y ya no en la escala de intervalo que suele suponerse por fiat.

Los instrumentos del AED (véase el capítulo 5) amplían enormemente el horizonte, siempre y cuando los resultados obtenidos de este modo no se utilicen en el mismo conjunto de datos para su confirmación. El análisis de los residuos y los valores atípicos mediante una apreciación gráfica permite una visión intuitiva de los resultados. Esto implica recalcular provisionalmente los modelos sin valores atípicos u otros datos potencialmente sesgados, e integrar y comprender el cambio resultante en los resultados. La comprensión se centra en qué cambia en los resultados si se omite exactamente tal o cual dato, cómo se relaciona esto con el algoritmo subyacente, etc.

7.8 Tamaños de los efectos

El cálculo de los tamaños del efecto debe hacerse dos veces, en la abstracción y en la escala original. Este último suele pasarse por alto. Por un lado, la *d* de Cohen, la *g* de Glass, etc. resultan ciertamente una buena elección cuando se trata de diferencias medias. Por otro lado, estas magnitudes resultantes no utilizan ninguna información nueva, ni siquiera en el caso de los modelos lineales (por ejemplo, η^2, f^2, etc.) o cuando se utilizan coeficientes de correlación *r*. Se limitan a presentar la información disponible de forma diferente y, por tanto, permiten la comparabilidad sin escalas entre estudios, como es necesario para los metaanálisis.La comparación de las diferencias o similitudes en las escalas originales (por ejemplo, las puntuaciones de los cuestionarios en los estudios psicológicos) obliga a abordar de nuevo la pregunta de investigación específica. Así, siempre hay que preguntarse: "¿Qué significa esta diferencia en la realidad?". Esta pregunta ya no se puede (¿afortunadamente?) responder tan sencillamente en abstracto. Más bien, es necesario evaluar concretamente qué efectos están presentes y cómo afectan a las personas, a los grupos o al comportamiento, al pensamiento, al sentimiento, a la acción, a la motivación, etc. para encontrar las respuestas adecuadas. En casos concretos, uno se dará cuenta de que tales afirmaciones anclada en el contenido son muy difíciles de formular. Entonces se necesita algo más que la mera afirmación estadística. Esto debe reflejarse y documentarse en consecuencia. La dirección y la cantidad de los efectos (errores de tipo S y de tipo M, véase el capítulo 4.3.3.2) deben debatirse a continuación con referencia al contenido.

7.9 Simulación

La validación o contrastación de las estimaciones de los parámetros mediante simulación (por ejemplo, bootstrap paramétrico, comprobaciones predictivas posteriores, etc.) examina la solidez de los modelos hallados. Entre otras cosas, esto produce intervalos de confianza para los parámetros y se amplía este espacio de la discusión del contenido. Se pueden examinar rangos enteros y no sólo coeficientes singulares. Si no se dispone de una simulación, los respectivos intervalos de confianza pueden calcularse igualmente a partir

de los errores estándar. Sin embargo, ni siquiera los intervalos de confianza son un remedio sólo porque no se utilice la significación, como se ha justificado anteriormente (véase el capítulo 4.3.5). Las simulaciones no pueden sustituir a las réplicas que faltan, pero pueden poner a prueba ciertos límites en el marco de la aleatoriedad generada de este modo.

7.10 Replicación – Replicación – Replicación

Nótese, sin embargo, el consejo sobre la replicación cuidadosa en las discusiones de Schimmack y Brunner (2017a), Nosek y Lakens (2014) y siguiendo a Marsman, Schönbrodt, Morey, Yao, Gelman y Wagenmakers (2017), así como en muchos otros autores. Al final de un proyecto de investigación, surgen naturalmente dos cuestiones importantes, cada una de las cuales abordaremos en una digresión. con más detalle en una digresión.

- ¿Son los resultados sólidos y reproducibles en diferentes contextos?
- ¿Ha merecido la pena el esfuerzo invertido (dinero, personas, tiempo)?

Sin mayor sorpresa, la primera pregunta ya conduce a la replicación, la segunda al cálculo completo de los costes.

7.10.1 ¿Por qué la replicación?

La tantas veces mencionada replicación de estudios tiene que ver principalmente con el proceso de conocimiento científico y no sólo con la estadística. Fisher (1935/1973, p.14) ya abogaba por la replicación:

> „To demonstrate that a natural phenomenon is experimentally demonstrable, we need not an isolated record, but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable whenwe knowhow to conduct an experiment [that] will rarely fail to give us a statistically significant result."

En relación con la prueba de significación, podemos decir que un fenómeno es demostrable experimentalmente cuando sabemos realizar un experimento [que] rara vez dejará de darnos un resultado estadísticamente significativo.

Desde el punto de vista de la teoría de la ciencia, la replicación se puede clasificar del siguiente modo: Así, según *la concepción estructuralista de la ciencia (non-statement view*, el concepto avanzado por Joseph D. Sneed, Wolfgang Stegmüller y otros), una teoría no es una entidad que pueda falsarse de algún modo, sino una red jerarquizada con un modelo matemático como núcleo estructural sin referencia empírica directa, del que emergen cada una de las leyes especiales de los elementos de la teoría (= proposiciones empíricas). La teoría de conjuntos se considera el lenguaje universal del empirismo en el concepto de teoría estructuralista. Es posible, mediante la aplicación coherente del principio de falsación, rechazar aplicaciones especiales de la teoría (pero no la teoría en sí) por ser científicamente insostenibles. Los intentos de aplicación exitosos y los fracasados responden en su conjunto a la cuestión de una operacionalización adecuada. Esto muestra una cierta cercanía a los programas de investigación de Imre Lakatos. Sin embargo, el concepto de teoría estructuralista es capaz de describir la relación de los elementos entre sí – mediante la teoría de conjuntos -- en una construcción teórica de forma mucho más precisa de lo que permite el programa de investigación de Lakatos. Los cambios en la construcción teórica de Lakatos también se formulan de forma más inespecífica y menos precisa. Se puede obtener más información sobre la función de la hipótesis en el proceso de

investigación, sobre la evaluación de la coherencia y operacionalizabilidad de las hipótesis y sobre las relaciones entre las hipótesis sustantivas y las formulaciones estadísticas (incluidas las predicciones), en los libros de Erdfelder y Bredenkamp (1994) y Hussy y Möller (1994).

La replicación de estudios es necesaria para realizar este trabajo teórico en la periferia. Un estudio no replicado podría ser un resultado casual. Sin embargo, si se encuentra el mismo efecto en diferentes muestras y se demuestra que es constante a las influencias contextuales, una base sólida lleva los efectos encontrados y las conclusiones resultantes. Se pueden estudiar los efectos tanto en el laboratorio como en la naturaleza, y se puede aprender mucho de los resultados potencialmente diferentes. A la hora de replicar, la cuestión es si se debe replicar algo en condiciones idénticas – y así "no aprender" de los errores del primer estudio – o en condiciones mejoradas. En ese caso, sin embargo, ya no se trataría de una réplica exacta, pero se seguiría investigando la constancia del fenómeno investigado en condiciones alteradas y contribuiría a demostrar la solidez de los posibles efectos. En cada caso concreto deberá aclararse dónde están los límites y un estudio sigue representando una réplica. En teoría, sin embargo, existe el peligro de que de repente se esté investigando otra cosa y ya no el fenómeno que se está analizando. Se deben reflexionar estos dos extremos ("no aprender nada" frente a "investigar algo completamente distinto") cuidadosamente.

Por desgracia, replicar estudios está bastante mal visto en ciencia, donde prevalece la idea de que "lo que se publica debe ser siempre nuevo y no debe conocerse todavía": al fin y al cabo, la estadística clásica busca sobre todo sucesos raros, ya que no puede decir nada sobre nada más. En consecuencia, los estudios de replicación se realizaban y se realizan muy raramente y aún más raramente – si es que alguna vez se publican – en revistas de alto nivel, como demuestran los intentos fallidos de replicación del estudio Bem (véase el capítulo 4.4.2.2). Esto contradice por completo el enfoque falsacionista de Popper – la comprobación crítica de las teorías y el proceso evolutivo de comprobación de las teorías. Ulrich, Erdfelder, Deutsch, Strauß, Brüggemann, Hannover, Tuschen-Caer, Kirschbaum, Blickle, Möller y Rief (2016) publicaron un importante artículo sobre este tema en el "Psychologische Rundschau", el órgano de la asociación de los psicólogos científicamente activos en Alemania (DGPS), que aborda el problema de los falsos positivos y, al mismo tiempo, la falta de réplicas y pide explícitamente estas últimas (véase también Erdfelder y Ullrich, 2018; Fiedler, 2018). Queda por ver si este impulso se reflejará finalmente en la comunidad científica y en las revistas pertinentes.

Dado que la estadística clásica presupone un muestreo aleatorio, en realidad no hay más remedio que replicar cada estudio para comprobar si los resultados tienen efectos constantes. Según la estadística clásica, sólo al aumentar el tamaño de la muestra se anulan mutuamente los errores. Otro problema es que, según la estadística clásica, no existen conocimientos previos que puedan aplicarse y utilizarse. Las muestras son efectos aleatorios, lo que significa que no pueden ampliarse sin más. Esto raya entonces en el "p-hacking", ampliando una muestra hasta que un efecto deseado alcanza un nivel "significativo". Tschirk (2014, p.80s.) muestra que este procedimiento puede conducir a resultados absurdos en la estadística clásica.

Este problema no existe en la estadística bayesiana. Sin embargo, la replicación también es igual de importante en este caso. En principio, se pueden utilizar los conocimientos previos existentes y acotar el área investigada en el sentido de una expectativa justificada empíricamente. En cierto modo, esto se corresponde con el concepto de falsificación, ya que si los datos empíricos se desvían fuertemente de la Prior, ésta experimenta un cierto cambio y, por tanto, falsación en el sentido de contraste de la realidad. No es necesario pretender que con cada nueva repetición surjan de nuevo la ingenuidad y la ignorancia y no se sepa nada, lo que por desgracia – poco sabiamente – constituye la base de la estadística clásica objetiva. Además, se pueden recoger datos hasta que los efectos estudiados se estabilicen. La estabilidad, sin embargo, es diferente de la significación.

7.10.2 Contabilidad de costes totales

Desde un punto de vista económico, nos preguntamos finalmente por el esfuerzo que supone encontrar un efecto de cierta magnitud y si éste tiene consecuencias reales para la práctica, la investigación posterior, la

sociedad, los (sub)grupos o los individuos. Esto nos lleva a una contabilidad de costes. Por desgracia, nadie lo hace realmente y publica el resultado al final de un artículo. Aunque no queremos restringir la libertad del investigador para investigar lo que le parezca subjetivamente interesante, seguimos considerando relevante la cuestión del beneficio público. A la inversa, esto crea naturalmente una enorme presión, ya que en el caso de la financiación por terceros, el presunto beneficio puede utilizarse indebidamente como medio de presión para demostrar el sentido y la finalidad de una determinada investigación o simplemente para conseguir que se apruebe una propuesta de investigación o, por el contrario, que se rechace. Tampoco queremos pensar en las ideas creativas que se les pueden ocurrir a algunas personas para justificar el beneficio incondicional de su propio trabajo. Esto nos lleva directamente a la ética en la ciencia, donde el autocompromiso definitivamente no es suficiente.

A pesar de estos obstáculos evidentes, la cuestión nos parece pertinente. Existen demasiados resultados científicos en paralelo, lo que hace cada vez más difícil separar el grano de la paja. Si nos preguntamos "¿Vale esto la pena en la práctica?" o "¿Es realmente un descubrimiento teóricamente significativo?", no debemos rehuir considerar precisamente efectos que no se han descubierto como una aportación valiosa a la ciencia. A veces, no encontrar algo o encontrar algo a pequeña escala puede ser más significativo que encontrar algo grande pero *trivial*. De hecho, esos resultados son más útiles porque se aprende más de los errores, corren el riesgo de fracasar y, en consecuencia, se exponen al riesgo de falsificación.

Los investigadores deberían adquirir el hábito de pensar económicamente, al menos internamente para sí mismos, lo que no significa que la investigación deba ser puramente lucrativa. Al contrario, la investigación puede y debe permitirse explorar fenómenos para los que no exista ya una apropiación económica indebida. El dinero invertido no determina si algo es útil. Pero la contabilidad de costes completos puede permitir hacerse una idea de si se repetiría un estudio a la vista de los resultados encontrados. El significado y el propósito y la contabilidad de costes totales son probablemente dialécticos entre sí y deben integrarse.

7.11 Límites del análisis de datos - la traducción interminable

7.11.1 Orientación al contenido y procesos de traducción

Dos aspectos saltan a la vista al final de la discusión sobre la estadística: la *orientación al contenido* y los *procesos de traducción* más o menos implícitos que tienen lugar durante el proceso de investigación. La perspectiva de contenido no debe descuidarse en la estadística, porque ésta es una secuencia de procesos de traducción. Se puede decir lo mismo, por cierto, del análisis cualitativo de datos. Las preguntas se traducen en modelos, los modelos en operacionalizaciones y variables relevantes. Éstas, a su vez, tienen que traducirse en ítems o indicadores, el grupo objetivo teórico tiene que traducirse en una muestra real, etc. – con los siguientes equivalentes en recogida de datos, evaluación, interpretación y aplicación. Los análisis cuantitativos sólo son cuantitativos y numéricos a corto plazo, porque gran parte del análisis tiene lugar como un complejo proceso de traducción y a menudo no es realmente ni una cosa ni otra.

Recordatorio 7.1: Procesos de traducción

La investigación es una serie interminable de procesos de traducción.

Como en todos los procesos de traducción – por ejemplo, partiendo de supuestos teóricos para llegar a relatividades empíricas y viceversa (Gigerenzer, 1981; Gürtler, 2005) – en las interfaces se pueden producir y se producen errores, errores de traducción e imprecisiones. O se producen grados de libertad inesperados porque no hay una transformación que preserve la identidad. Todo esto hay que descubrirlo y, si es

necesario, corregirlo o evaluar y documentar la inexactitud. Por lo tanto, hay que examinar la plausibilidad del contenido de los resultados. Si esto no es posible, se introduce una incertidumbre adicional y la visión de conjunto se distorsiona un poco más. A la inversa, esto puede ser una oportunidad porque, sobre el trasfondo de análisis sólidos, podemos utilizar creativamente todos nuestros conocimientos para tomar decisiones significativas y plausibles en un mundo no determinado, basándonos en toda la información disponible. Nuestra información contiene mucho más que análisis numéricos y semánticos. Se convierte en un problema cuando nos quedamos atascados en la información abstracta y, en el caso de los análisis cuantitativos, en forma de resultados numéricos, y no podemos clasificarla como relativa. La interpretación de los datos como tal es un acto cualitativo, no numérico. Como deberían haber demostrado las observaciones sobre estadística, en estadística se toman muchas decisiones que no se pueden justificar únicamente con coeficientes y gráficos. Se requiere un compromiso cualitativo con los resultados. Un caso extremo sería la cuestión de los límites de significación, en la que, en última instancia, el sentido común se ha externalizado por completo a una convención sin motivo y sigue sin justificarse ni cuestionarse.

Conceptos como el ROPE reintroducen el sentido común porque a lo largo de un intervalo se deben examinar los datos para distinguir la significación de la insignificancia en las escalas originales. Especialmente en los metaestudios, el reto es no perderse en los coeficientes, sino mantener la referencia contextual en cada caso, a pesar de los diferentes contextos que entran en el metaanálisis. Los métodos mixtos en el sentido de traducción continua siempre forman parte implícitamente del proceso, pero a menudo esto no se entiende o ni siquiera se hace explícito. Esto se hace más difícil cuando los artículos empiezan a perder cada vez más componentes del contenido de los estudios e informan cada vez menos. El contenido experimenta su límite natural en el contexto de los meta-estudios, pero a pesar de ello es esencial una cantidad mínima de información.

> Caso 7.1: Estadística y realidad
>
> Por ejemplo, podemos demostrar que dos grupos son estadísticamente diferentes entre sí o que varias variables tienen una relación lineal estadísticamente significativa entre sí. Sin embargo, la estadística no puede decir nada sobre lo significativo que es ese resultado en la realidad y cómo se produce en términos de contenido. A ello no ayudan los tamaños del efecto, que por definición carecen de unidad, sino únicamente la integración dialéctica de las teorías y modelos, datos empíricos y resultados, y una discusión de los datos en su escala original.

Pero eso no es todo: por desgracia, se pueden publicar los estudios cuantitativos también completamente sin contenido, lo que consideramos peligroso. Un ejemplo de este tipo de estudios con meta-carácter, donde el contenido ya no es reconocible, es el gran estudio de replicación de Camerer, Dreber, Holzmeister, Ho, Huber, Johannesson, Kirchler, Nave, Nosek y Pfeier (2018). En este gran estudio, que equivale a un meta-estudio en términos de diseño y ejecución, se replicaron sistemáticamente 21 estudios experimentales seleccionados de las ciencias sociales que se publicaron en la revista *Nature and Science* entre 2010 y 2015. Además del altísimo esfuerzo por replicar los estudios cuidadosamente y en red con otros investigadores, sin embargo, el aspecto negativo del artículo es que desde el principio hasta el final no tienes ni idea de lo que realmente estás leyendo. Por supuesto, sabes que se trata de experimentos de ciencias sociales, nada más. Esto no sólo es poco atractivo para el lector, sino que también tiene consecuencias para el contenido. En el artículo escrito con eficacia, hay una discusión que está completamente desvinculada del contenido y exclusivamente abstracto-metodológico sobre el tema de las réplicas, tamaños del efecto, significancias, factores de Bayes, etc. – no se menciona en absoluto lo que tratan en detalle los estudios que fueron replicados. La metodología se convierte en un fin en sí mismo, lo que probablemente es incluso pretendido aquí. En este sentido, habría que examinar más detenidamente los estudios individuales o al menos los resúmenes para ver si la agrupación de tamaños de efecto abstractos a nivel meta-analítico tiene (o no) sentido en algún momento. Los coeficientes y los gráficos de los resultados comunicados no bastan por sí

solos para comprenderlo. Que se pueda hacer algo no significa que esté justificado en cuanto al fondo. Incluso se omite una breve enumeración tabular de los sujetos, el diseño y la muestra de cada estudio, aunque todos ellos están prolijamente referenciados y enumerados. Por un lado, es comprensible que las revistas intenten reducir al mínimo los artículos – y éste es obviamente un artículo metodológico. Sin embargo, de nada sirve saber si los estudios son replicables o no si ya no se tiene ni idea de lo que se habla y se juzga. El aspecto del contenido es intemporalmente relevante, porque puede ser que ciertas características del contenido de los estudios (tema, diseño, muestra, ...) aumenten o disminuyan el éxito de los intentos de replicación y esta discusión no se lleva a cabo en el artículo. Los coeficientes no lo justifican tan fácilmente. Se omite la vinculación y la influencia mutua del objeto de investigación y la metodología o el análisis de datos. Una comparación abstracta trata entonces todos los estudios como equivalentes en términos de replicabilidad, pero esto puede no ser el caso por razones sustantivas. Dado que este debate no se celebra, no lo sabemos.

7.11.2 Garantía de calidad

Otro ejemplo es la *garantía de calidad*. Por ejemplo, basándonos en análisis estadísticos comprobamos que la calidad de los productos de una producción está sujeta a fluctuaciones más o menos mínimas. Dejamos a un lado las razones, los problemas pueden residir en la calidad de las materias primas, en la atención cambiante del personal o en irregularidades relacionadas con las máquinas o son simplemente causadas por la forma del día ("calidad de lunes"). Pero, ¿en qué momento detenemos la producción del día y pasamos a la resolución de problemas? Una parada de este tipo significa que la producción no puede continuar y esto supone costes y esfuerzos elevados (personal, resolución de problemas, pérdida de ingresos, efectos sobre clientes y proveedores, etc.). En este sentido, el proceso de decisión es muy importante y sólo debe llevar a detener realmente la producción bajo las indicaciones más claras posibles. Así pues, ahora nos damos cuenta de que la calidad de un producto fluctúa en una jornada laboral determinada. Pero, ¿son estas fluctuaciones suficientes para detener la producción? ¿Qué importancia tienen? Las estadísticas por sí solas no pueden responder a esta pregunta. Hay que basarse en criterios objetivos (por ejemplo, mediciones físicas, información previa, etc.), valores empíricos (estas fluctuaciones pueden o no suponer ninguna diferencia en la práctica) o consideraciones de coste-beneficio (¿es más cara la parada de la producción que las consecuencias de la proliferación de productos potencialmente defectuosos). En el caso de los productos tecnológicos de alto rendimiento (por ejemplo, la informática, la aeronáutica y la fabricación de automóviles, véase el estudio de caso sobre el enfoque Neyman-Pearson en el capítulo 4.3.3.4), incluso fluctuaciones mínimas pueden poner en peligro la seguridad hasta tal punto que una parada de la producción se hace indispensable porque los riesgos potenciales para la vida y la integridad física son demasiado grandes y esto orienta masivamente las consideraciones de coste-beneficio en una dirección determinada. Las estadísticas adquieren entonces una importante función indicativa, ya que ayudan a proporcionar información importante al margen de subjetividades y peritajes. Señala diferencias y correlaciones. Pero éstas deben reunirse de forma coherente y plausible para formar una decisión. La decisión queda en manos del ser humano, aunque en general hay cada vez más tendencias (por ejemplo, diagnósticos médicos, robots militares, sistemas de alerta general) a externalizar las decisiones al campo de la tecnología informática y la inteligencia artificial (IA).

7.11.3 Automatización

Por supuesto, son concebibles escenarios en los que los mecanismos automatizados de toma de decisiones por parte de la IA resulten útiles. Sin embargo, las decisiones tienen consecuencias, ya sea en los drones armados, en los seguros para evaluar a los clientes o en los coches de conducción autónoma. Pero incluso estos mecanismos automáticos no son más que un mapeo de aquellos procesos y reglas de toma de decisiones que siguen siendo humanos.

Los algoritmos ya se utilizan para predecir la reincidencia en el sistema penal o por la policía a través de la minería de datos para predecir delitos en determinadas zonas (por ejemplo, robos). Sin embargo, todo esto no es un producto independiente y descontextualizado de la inteligencia artificial, sino una aplicación coherente de conocimientos cualitativos que co-determinan los algoritmos, que sólo pueden ejecutarse de forma automática. Aquí no queremos hablar todavía de inteligencia, sino de algoritmos bastante complejos que procesan datos de forma regida por reglas y determinan y modifican criterios en el proceso. Puede que la inteligencia real no sea posible en el ordenador puramente pero sólo cuando la combinación de chips informáticos y biotecnología dé algunos saltos más. En la IA actual vemos una integración totalmente satisfactoria de métodos cualitativos y cuantitativos, aunque estos métodos sigan siendo defectuosos y no puedan sustituir completamente a los humanos. Esto último porque no se trata de inteligencia en el sentido psicológico.

7.11.4 Reconocer la significación

La estadística siempre es útil cuando se trata de reducir las relaciones numéricas y llevarlas al punto y estimarlas o probarlas en relación con un criterio sustantivo. La estadística asume la función de indicación, haciendo afirmaciones sobre las relaciones de los datos y sobre cómo se pueden clasificar los resultados en términos probabilísticos o qué límites de tolerancia obtenemos en los resultados. La estadística, sin embargo, no puede decidir por sí misma, sino que el resultado de un análisis estadístico debe (re)traducirse al contexto situacional del que proceden los datos para llegar a una decisión plausible. En este contexto, se debe examinar si un resultado obtenido es pertinente y significativo para la acción o no. Esta decisión se plantea sobre una base cualitativa y sustantiva. Su naturaleza no es cuantitativa, aunque la decisión se base en un análisis cuantitativo y se pueda automatizarla.

¿Qué importancia tiene una diferencia o una conexión? Imaginemos que en una larga serie de pruebas infructuosas de un nuevo medicamento en la investigación clínica farmacéutica, un paciente parece recuperarse. Estadísticamente, este único paciente no importa, pero en términos de contenido y a nivel práctico, este paciente podría ser la pista de un gran avance. Entonces, ¿prestamos atención al paciente o no? La estadística clásica podría afirmar que se trata de un valor atípico aleatorio o de un caso aislado irrelevante, porque el nuevo fármaco no ayudó a todos los demás pacientes. Sin embargo, la intuición clínica exige que nos fijemos muy bien en por qué este paciente está mejorando y qué factores intervienen. Esto puede conducir a la mejora de un fármaco o un tratamiento y a que las diferencias sean detectables posteriormente en (sub)grupos o muestras más grandes. En consecuencia, la estadística no puede sustituir al análisis de casos únicos y contados. Si en tal caso se aplican ciegamente las convenciones – lo que corresponde a "más de lo mismo" – puede que no se consiga el necesario logro transformador de descubrir lo nuevo. Pero también se puede dar el caso de que sólo haya que seguir un procedimiento estándar (Cecchin, Lane & Ray, 2002) para descubrir algo nuevo.

Y esta diferenciación – "¿Cuál es el caso?" – hace de la investigación un problema complejo, ya que va mucho más allá de la definición de una simple tarea. Esto aboga claramente en contra de una aplicación ciega de los procedimientos estadísticos clásicos de NHST en sentido amplio, que no funcionan (no pueden funcionar) en absoluto de forma adecuada al caso y a menudo necesitan demasiadas condiciones previas para para ser utilizados. Se deben aclarar las cuestiones de significación antes de realizar una investigación. A menudo, sin embargo, esto no es posible porque se dispone de muy poca experiencia y supuestos previos para hacer predicciones razonables sobre los límites de la significatividad. Algunas cosas se pueden poner de manifiesto sólo a través del compromiso directo con el tema en el propio proceso de investigación. Otra razón es que en muchas ramas de la investigación se evita la repetición generalizada en lugar de producir constantemente piezas nuevas y más o menos inconexas del rompecabezas. Esto suele acabar en *la menor unidad publicable*. El resultado es una masa de publicaciones difícil de gestionar.

Por ejemplo, en la preparación de un estudio sobre el uso de un entorno de aprendizaje innovador en las escuelas, podemos preguntarnos si una mejora de 0.2 a 0.4 puntos en una escala Likert sobre la autopercepción de los alumnos es lo suficientemente significativa como para justificar el uso (personal,

material, finanzas y otros recursos) de un programa intensivo de formación de profesores de un año de duración con intervenciones de acompañamiento a nivel escolar a gran escala. Y entonces podemos preguntarnos si merece la pena esforzarse por integrar exactamente esta forma de trabajar en la formación general del profesorado. También podemos preguntarnos si nos merece la pena dedicar toda nuestra vida a un estudio empírico tan elaborado. La tercera parte financiadora probablemente se esté haciendo la misma pregunta, sólo que a un nivel monetario de porcentajes de trabajo y recursos materiales.

Podríamos formular la misma pregunta de otro modo, preguntando a expertos en la materia antes de realizar este estudio. Entonces preguntaríamos qué cambios específicos esperan ver en las herramientas de autoconciencia que se utilizarán cuando la formación y el entrenamiento de los profesores se llevan a cabo según todas las reglas del arte (Wahl, 2006) y los instrumentos de medición utilizados son conocidos por todos. Esto es exactamente lo que hemos hecho en un estudio de investigación educativa empírica sobre una base cualitativa informal entre colegas, todos ellos al menos habilitados o profesores en el campo de la psicología. Curiosamente, o tal vez incluso atemorizadoramente, ninguno de los expertos entrevistados pudo dar ninguna afirmación clara en absoluto (sobre este fenómeno Brandstätter, 1999). La teoría, los estudios preliminares disponibles y mucho más eran tan heterogéneos que sencillamente no era posible hacer una afirmación clara y vinculante. Pero eso no bastaba. Parece mucho más aterrador que aparentemente no sea posible para los profesores situar el resultado de una investigación en un contexto real, es decir, situar dichos cambios en una escala de Likert en hechos reales en el aula, en la autopercepción de los niños, etc. El consenso de diferentes expertos sobre los resultados de la investigación no fue suficiente. Ni siquiera queremos hablar de un consenso de diferentes expertos sobre esta cuestión – aparte de su inseguridad, expresada con gran autenticidad, de no poder decir nada concreto. No queremos desvalorizar a los colegas entrevistados aquí, sino todo lo contrario. Nos pareció extremadamente profesional que expresaran su propia inseguridad exactamente de esta manera. Sólo por *razones formales* mencionamos que, por supuesto, *nosotros mismos no podemos hacerlo* y no sabemos realmente cómo evaluar mejor este tipo de situaciones, a pesar de haberlas tratado intensamente.

Creemos que aquí sale a la luz un grave déficit general de la ciencia: la interfaz de la ciencia con la realidad cuando se trata de la utilización y aplicación de resultados empíricos, cuando el tema es la evaluación del empirismo del contexto científico para la práctica cotidiana y cómo sincronizar con precisión ambos niveles: ciencia y práctica. Para nosotros, esto significa que, en el contexto de los métodos mixtos, no sólo hay que invertir en las estadísticas, sino sobre todo en los *procesos de traducción* entre las fases del proceso científico y la realidad.

Parte III

Métodos Cualitativos

Capítulo 8

¿Qué es lo Cualitativo en Realidad?

"Sin especulación no hay nueva observación".

Carta a A.R. Whitehead, 22.12.1857
Charles Darwin, 1809-1882

8.1 Preludio

El término cualitativo se remonta al latín *qualitativus*. Denota la bondad o cualidad de algo o alguien y pretende expresar características típicas. La cantidad o cuantitativo, es decir, el número o montón de algo, suele figurar como opuesto o contrapalabra. Se trata de una figura dialéctica: aunque superficialmente debería destacar lo cualitativo, es decir, la calidad del objeto en cuestión, esto no suele bastarnos a los humanos. Siempre queremos *más*, y normalmente *más de lo mismo*. La calidad requiere necesariamente una cantidad mínima, de lo contrario es una cantidad nula y poco útil para nosotros. Lo que es excelente pero no está, sigue sin estar. A la inversa, la cantidad requiere un mínimo de calidad, de lo contrario tampoco nos sirve. Si algo es común pero no es bueno en absoluto, nos sirve de poco. No en vano existe el dicho "quien compra barato, compra dos veces". La cuestión ahora es: ¿existen ambas cosas a la vez? ¿Calidad *y* cantidad? En principio, no hay nada que objetar, pero la práctica demuestra que no es tan sencillo. Podríamos preguntarnos del mismo modo, ¿en qué unidades se mide una calidad? En el lenguaje cotidiano acabamos hablando de calidad "baja", "media" o "alta". Así pues, la calidad suele recibir un término cuantitativo, para especificarla con mayor precisión. De este modo, la calidad adquiere su significado por comparación con una cantidad comparativa que no se tiene necesariamente que especificar con más detalle. Apartándonos sólo superficialmente de esto, encontramos afirmaciones como que algo es "cualitativamente diferente" y precisamente *no* un "más" o un "menos" son adecuadas para describir o reconstruir adecuadamente y, en última instancia, comprender esta cualidad.

La cualidad se destaca de forma inespecífica, lo que no excluye la introducción de una especificación cuantitativa en el caso individual concreto. Si una cualidad destaca por sí misma, la probabilidad de que sea necesaria una especificación cuantitativa parece menor. Si, por el contrario, se añade un nivel de comparación con otra cualidad de cualquier tipo, entonces en nuestra opinión, aumenta la probabilidad de que vaya acompañada de una evaluación cuantitativa. Utilizamos deliberadamente el concepto de probabilidad como modelo para describir los cambios descritos al tratar una calidad.

Otto Dempwol (1939, p.26) define lo cualitativo yendo más allá de la abstracción del lenguaje, las ideas y los sentidos y extrayendo de ellos lo particular:

> "§41. Es un artificio del pensamiento humano separar una parte de las impresiones sensoriales, que permanecen iguales en el tiempo, y desprenderla del conjunto de las cosas como un grupo imaginativo especial de cualidades. La expresión lingüística para esto es la palabra tipo cualitativo (=adjetivo)".

¿Todo esto, es decir, la posible mezcla de calidad y cantidad, va en detrimento de la calidad per se? Nosotros responderíamos: ¡No, desde luego que no! Una especificación cuantitativa es una especificación dentro de un contexto concreto en el que, obviamente, las comparaciones desempeñan un papel. El hecho de que estas comparaciones sean o no significativas no tiene nada que ver con ello y sería objeto de una discusión aparte. Pero si son significativas, deberíamos utilizar este nivel. Si comparamos los colores rojo y verde, inicialmente tenemos a mano dos colores diferentes, por lo que no podemos decir "mejor" o "peor". Por supuesto, podríamos pedir preferencias y algunos preferirían el rojo y otros el verde. No queremos llegar ahí, porque esto no tiene que ver con la calidad en sí, sino sólo con la percepción de la calidad por parte de los destinatarios. Sin embargo, si examinamos qué es realmente el rojo o el verde, nos encontramos rápidamente con el espectro cromático y, en términos de colores espectrales, el rojo y el verde pueden traducirse en términos cuantitativos. El rojo se sitúa entre ≈700 y 630 nm (longitud de onda) y ≈430-480 THz (gama de frecuencias) y el verde entre ≈560-490 nm o ≈540-610 THz. Podríamos decir que el espectro del rojo está por encima del del verde o que el espectro del verde está por debajo del del rojo; y ahí hemos cuantificado. Sin embargo, esto no dice nada sobre si el rojo es ahora mejor, igual o peor que el verde. Para eso necesitamos un contexto: por ejemplo, que cuanto más cerca del espectro azul y ultravioleta está la luz, más perjudicial es para la retina a medida que aumenta la energía. La calidad del rojo y del verde, así como la belleza del sol de la mañana y de la puesta de sol, no se ven afectadas por la posibilidad de cuantificación, mientras podamos disfrutar del sol de la mañana y de la puesta de sol, del césped verde y de los bosques, o de las rosas rojas y las amapolas, sin más. La calidad no disminuye por el hecho de que sepamos cómo se produce ni aumenta especialmente. La percepción subjetiva de la calidad – quedémonos con el amanecer – no cambia realmente. Si cambia, se debe de nuevo a la reacción del observador, que puede no ser capaz de manejar esta información. Pero incluso esta reacción podríamos describir bien por una conexión dialéctica de *más*/*menos* o *diferente*/*igual*. Otro ejemplo: un argumento en una discusión puede ser bueno o malo, lo que indirectamente significa que su calidad es alta o baja. Y ya, con los términos alto y bajo, tenemos de nuevo en juego el nivel cuantitativo. Hay muchos otros ejemplos. Y es difícil mantenerse al margen y a veces ni siquiera se desea.

A lo que evidentemente queremos llegar es a que la estricta separación entre calidad (= *métodos cualitativos de análisis de datos*) y cantidad (= *métodos estadísticos de análisis de datos*), en la que se hace hincapié sobre todo en las ciencias sociales, sólo existe en la superficie. Si profundizamos un poco más en respectivos temas, ambos niveles se entremezclan y uno se funde perfectamente en el otro. Incluso diríamos que uno no puede existir sin el otro. Muy a menudo los científicos utilizan un nivel para describir el otro, pero sin ser necesariamente conscientes de ello (Gürtler & Huber, 2006) y aunque (o precisamente porque) rara vez ambos son el centro de interés al mismo tiempo con la misma pretensión. Consideramos que la dialéctica es un modelo adecuado para integrar los opuestos, a menudo sólo superficiales, de calidad y cantidad de forma adecuada al caso. Esto no excluye que haya situaciones en las que uno de ellos pase a primer plano con tanta fuerza que el otro parezca entonces superfluo. Pero quizá no. Tal caso depende de la cuestión. Casos ejemplares serían los análisis que a menudo se llevan a cabo en el contexto de la hermenéutica objetiva (véase el capítulo 11) y en los que realmente vemos problemas para incluir elegantemente un nivel cuantitativo. A la inversa, sin duda hay preguntas comparables que pueden responderse con números tan bien que no necesitamos preguntarnos por el nivel subjetivo o intersubjetivo del significado y sus implicaciones.

8.2 Métodos cualitativos

En este contexto, ¿qué son los métodos cualitativos? En el procedimiento de exclusión, podemos definir que se trata aquí de todos los *métodos de recogida y análisis de datos no numéricos* y continuar coherentemente con la definición conceptual anterior de calidad y cantidad. Formulados positivamente los métodos cualitativos sirven para recoger y procesar datos (en su mayoría) textuales en el sentido más amplio. Cuando hablamos aquí de *textos*, los entendemos con Oevermann, Allert, Konau y Krambeck (1979, p.378; véanse

los capítulos 11.3 y 11.4) como la *documentación (protocolos) de acciones e interacciones sociales, independientemente del tipo de medio*. Todo lo que pueda archivarse está permitido. En concreto, pues, se trata de obtener información para responder a las preguntas de la investigación con ayuda de entrevistas, registros de observación, notas de campo, materiales de archivo, publicaciones en todo tipo de medios, transcripciones de discursos, grabaciones sonoras, fotos, vídeos, etc. y analizarlos de manera que encontremos en ellos respuestas a nuestras preguntas de investigación. ¿Se trata de datos cualitativos o cuantitativos? ¿Son palabras que adquieren significado o frecuencias que cambian?

Los numerosos métodos de *recogida de datos* cualitativos no pueden ser, por supuesto, objeto de este libro. Introducciones generales y temas en profundidad sobre la investigación social cualitativa son ofrecidos por Glaser y Strauss (1979), Flick, Kardo y Steinke (2000), Mayring (1990) y muchos otros autores. Entre otras cosas, tratan de métodos generales para obtener datos cualitativos, enfoques relevantes como la teoría fundamentada o temas en profundidad como la triangulación, los criterios de calidad, la tipificación o la forma de extraer conclusiones. Aquí nos centramos en cómo *analizar* los datos recogidos en entrevistas, archivos etc.

¿Para qué sirven los métodos cualitativos? También podemos responder a esta pregunta descartando primero las alternativas: la investigación de nuevas cuestiones se encuentra todavía en un terreno más o menos desconocido y suele faltar orientación en forma de teorías o hipótesis. Por lo tanto, no se puede recurrir simplemente a instrumentos ya probados, como pruebas o esquemas de observación. Al adentrarse en territorio desconocido, primero hay que intentar orientarse para construir hipótesis adecuadas. Camic, Rhodes y Yardley (2003, págs. 8-12) dan respuestas claramente positivas a la pregunta sobre el significado y la finalidad de los métodos cualitativos: *Los métodos cualitativos ayudan a explorar áreas problemáticas y a elaborar hipótesis o teorías*. Además, los métodos cualitativos permiten realizar análisis *contextualizados* o *situados* de los acontecimientos. Además, permiten identificar características de fenómenos complejos dinámicos o inusuales para los que no se dispone de otros métodos precisamente por estas características. Esto es especialmente cierto en el caso del acceso al mundo subjetivo de los participantes en la investigación, por ejemplo, a sus convicciones, creencias y los significados que atribuyen a los fenómenos. Los autores se refieren en particular a la dimensión estética de la experiencia como un ámbito de investigación en el que los investigadores dependen de los enfoques cualitativos. Además, a diferencia de las pruebas estandarizadas, los métodos cualitativos pueden tener en cuenta las relaciones de los participantes en la investigación entre sí y con su contexto sociocultural y socio-interactivo más amplio. Por último, los métodos cualitativos implican una reflexión sobre dichas relaciones en el contexto de la investigación.

Además, los métodos cualitativos permiten crear las denominadas estructuras de casos (véase el capítulo 11.5.1), que tienen potencial no sólo para descripciones, sino también para explicaciones, predicciones e intervenciones en la práctica (por ejemplo, Studer, 1998; Gürtler, Studer & Scholz, 2012). A partir de las estructuras causa-efecto se pueden hacer deducciones sobre las relaciones causales y, en principio, se gana acceso a verificarlas. Esto sitúa a los métodos cualitativos al mismo nivel que los métodos cuantitativos de análisis de datos. Como ocurre a menudo, la calidad de la aplicación y ejecución en el caso concreto nos parece el criterio central de la cientificidad y no los métodos en sí.

De acuerdo con nuestro *modelo cíclico del proceso de investigación* (véase la fig. 2.1), para muchos autores *la pregunta de investigación* ya determina el uso de enfoques de investigación cuantitativos o cualitativos. Maxwell (1996) distingue entre preguntas instrumentalistas y realistas, por un lado, varianza por un lado, y preguntas orientadas a la varianza y preguntas orientadas al proceso, por otro. En perspectiva instrumentalista, la pregunta de investigación pide respuestas a lo que los participantes en un estudio informan directamente sobre un fenómeno o lo que se puede observar directamente, mientras que se pueden formular las preguntas desde la posición del realismo científico (Chakravartty, 2016) "en expresiones de comportamientos inferidos, estados psicológicos o influencias causales" (Maxwell, 1996, p.56).

Así, para un estudio de la situación de los alumnos superdotados, cabría preguntarse "¿Qué dicen los alumnos superdotados sobre su situación en su clase?" frente a "¿Cómo se sienten los alumnos superdotados en su clase?" En el primer caso, un documento que resuma y cuantifique lo que dicen los alumnos responde a la pregunta de investigación. En el segundo caso, se esperan interpretaciones de las declaraciones y conclusiones que apunten más allá de la frecuencia de las declaraciones individuales. Quizá sea aún más evidente la determinación del método por la pregunta al distinguir entre preguntas de varianza y de proceso.

Un ejemplo del primer tipo de pregunta sería "¿Quieren más niños o niñas superdotados que se les enseñe en clases especiales y por qué?", mientras que para el segundo tipo se podría preguntar, "¿Qué opinan los niños y niñas superdotados de que se les enseñe en clases especiales y por qué?" Ambas preguntas buscan explicaciones como respuestas, pero con el énfasis en el primer caso en las diferencias cuantitativas y los factores causales, en el segundo caso con énfasis en los aspectos cualitativos de cómo se producen las actitudes (posiblemente diferentes). Ambas cuestiones son legítimas en sí mismas y exploran aspectos diferentes. Sin embargo, tomadas en conjunto e integradas, pueden ir más allá de sus resultados individuales y proporcionar una imagen mucho más completa, exactamente en el sentido del teorema de la Gestalt.

Marecek (2003, p.57) destaca la importancia de los tipos de preguntas "*¿por qué?*" y "*¿cómo?*", señalando que la investigación cualitativa trata de "cómo la agencia y el significado humanos están constituidos por el flujo constante de la vida social y cultural". El *cómo* de la pregunta de investigación requiere métodos de investigación cualitativa

- que recogen datos de los mundos vitales de los implicados,
- que son vistos en su contexto sociocultural y
- que son vistos como sujetos reflexivos.

Esto último no es relevante desde el punto de vista de la Hermenéutica Objetiva (véase el capítulo 11), sino que "sólo" interesan los procesos objetivamente (es decir, intersubjetivamente) reconstruibles en la vida y las acciones de las personas. Creswell (2009, p.130) recomienda explícitamente la *pregunta del cómo* para los estudios cualitativos y la distingue del *por qué* de los enfoques cuantitativos:

> „Begin the research questions with the words what or how to convey an open and emerging design. The word why often implies that the researcher is trying to explain why something occurs, and this suggests to me a cause-and-e ect type of thinking that I associate with quantitative research instead of the more open and emerging stance of qualitative research.“

Una vez más, hay que citar la Hermenéutica Objetiva, que aborda tanto las cuestiones de "¿Qué se puede reconstruir? ¿Cuál es su impacto concreto? ¿Qué significado tiene? ¿Por qué alguien actúa como lo hace?" como preguntas legítimas. Vemos que hay mucha heterogeneidad en los enfoques cualitativos y que las demarcaciones del campo cuantitativo son ya difíciles de trazar en las preguntas de investigación, lo que no significa que no estén ahí. En sus observaciones posteriores, Cresswell (2009, p.130s.) sugiere el uso de verbos más exploratorios en el camino que va del planteamiento de un problema a la pregunta de investigación, porque transmiten el "diseño emergente" y, por lo tanto, vinculan la pregunta, la recopilación de datos y el análisis. Recomienda (ibíd.)

"Diga al lector que su estudio

- descubre algo (por ejemplo, la teoría fundamentada)
- intenta comprender algo (por ejemplo, la etnografía)
- investiga un proceso (por ejemplo, un estudio de caso)
- describe experiencias (por ejemplo, fenomenología)
- relata historias (por ejemplo, la investigación narrativa)".

Sin embargo, la tarea del análisis cualitativo de datos no es sólo la exploración, como ya hemos explicado anteriormente. Para nosotros, ningún tipo de análisis de datos, cuantitativo o cualitativo, es realmente superior al otro, ya que están integrados y deben complementarse. Además, responden a preguntas diferentes: ¿Cómo se pueden comparar de forma significativa? Parece más razonable plantearse cómo obtener y analizar la información pertinente y no qué paradigma y creencia se sigue, siempre que el procedimiento se aplique de forma metódicamente controlada y regida por normas, es decir, científico.

8.3 Estructura y objetivos de aprendizaje

Entre los muchos enfoques para analizar el significado de los datos no numéricos o cualitativos, distinguimos dos enfoques (véase la Fig. 8.1): Los procedimientos para *codificar unidades de significado* en los textos, es decir, para asignarlas a categorías, en resumen el *paradigma de codificación* (s. cap. 9), y los procedimientos para reconstruir estructuras de acción y significado a partir de los textos, en resumen, el *paradigma de la reconstrucción* (véase el capítulo 11). Aquí debemos distinguir de nuevo entre dos alternativas metodológicamente amplias:

1. Enfoques para el análisis del contenido manifiesto, por un lado, y
2. Enfoques de análisis del contenido latente de los datos, por otro.

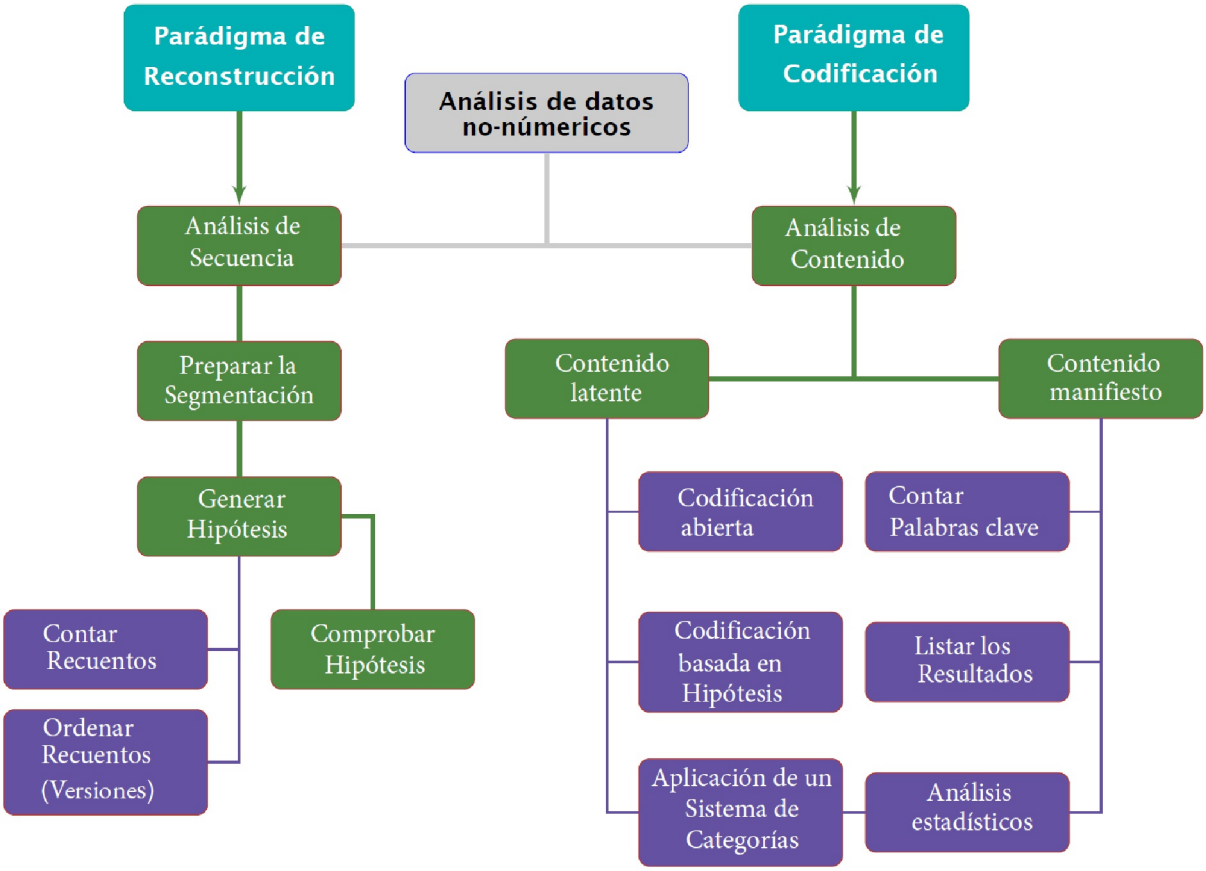


Figura 8.1 *Resumen de los procedimientos de análisis de datos cualitativos*

El paradigma de la codificación está mucho más extendido en la práctica de la investigación social cualitativa que el paradigma de la reconstrucción. Empezaremos por él en el próximo capítulo 9, examinaremos este enfoque en detalle y pasaremos a un ejemplo práctico (véase el capítulo 9.6). En realidad, el paradigma de reconstrucción no se encuentra en los países angloamericanos o de habla no alemana. El capítulo 11 trata de esta herramienta increíblemente potente y también concluye con el mismo estudio de caso en el capítulo 11.13, que ya se utiliza en el paradigma de codificación. Entre medias, el análisis cuantitativo de textos se inserta en el capítulo 10, también con el mismo estudio de caso (véase el capítulo

10.1). Esta repetición permite comparar la aplicación de los tres enfoques analíticos diferentes al mismo material de datos en cuanto a los resultados obtenidos.

El *paradigma de la codificación* se ha utilizado desde los inicios de la recepción por parte de las ciencias sociales de los procedimientos de análisis de textos, ya sea en forma de *análisis de contenido cualitativo o cuantitativo*. En 1952, Kracauer (1952) y Berelson (1952) publicaron artículos fundamentales. Mientras que para Kracauer el análisis de contenido cualitativo intenta revelar las categorías de significado ocultas y *latentes* del texto independientemente del contenido manifiesto del mismo, para Berelson el análisis de contenido cuantitativo sirve para registrar de forma sistemática, objetiva y cuantitativa el contenido *manifiesto* de la comunicación.

En el *análisis cualitativo de contenido* o de texto, siguiendo a Miles y Huberman (1984; 1994), se distinguen dos procesos interrelacionados, *la reducción del texto* y *la conclusión*. Dependiendo de la variante metodológica, se procede a la reducción según un procedimiento más o menos fuertemente estructurado. La característica común de estos procedimientos es la *clasificación o categorización* de las secciones del texto. En la práctica, esto significa reducción y condensación de información.

Siguiendo el uso lingüístico de Miles y Huberman (1984), nos referimos a los símbolos que sirven para identificar las categorías como *códigos*, y al proceso de reducción en consecuencia como *codificación* de los datos. Esto no es más que una reducción resumida del texto con categorías (códigos) adaptadas a la pregunta de investigación y al contenido. En el proceso de conclusión, se intenta a continuación encontrar configuraciones típicas y/o recurrentes entre estos códigos y elaborar sus relaciones entre sí. Además se puede prestar atención a los patrones de secuencia típicos de los códigos, buscar superposiciones y subordinaciones de códigos, combinar códigos con significados similares en una categoría superordinada ("metacódigo"), entender los códigos como polos de una dimensión común o formular hipótesis sobre las conexiones entre los segmentos de texto codificados y ponerlas a prueba específicamente.

El *análisis cuantitativo del contenido* o del texto se centra en las características manifiestas del texto, es decir determinadas palabras clave, expresiones idiomáticas, metáforas, etc. Se buscan y se cuentan. Lo que interesa son los elementos textuales accesibles y directamente determinables. La restricción al contenido manifiesto es, sin embargo, sólo aparentemente posible, porque la definición de elementos críticos del texto para el análisis cuantitativo de contenido implica necesariamente la suposición de que precisamente estos elementos directamente accesibles se refieren a significados del texto que sólo pueden averiguarse indirectamente. En otras palabras, antes del análisis cuantitativo, se deben definir los códigos en forma de palabras clave. Es necesario determinar las palabras clave como representantes del significado – como los códigos. Un recuento sin interpretaciones es imposible. En última instancia, también en el análisis de contenido cuantitativo se extraen conclusiones de un contenido manifiesto y observable a un contenido de significado latente. Esto requiere un conjunto de reglas claras y comprensibles sobre qué términos tienen qué significado (información) y cómo inferir de ello cualquier latencia.

El *análisis de secuencias* como *la forma prototípica* de aplicación del *paradigma de reconstrucción* no parte de una visión general del texto en su conjunto para buscar segmentos de texto que sean relevante para la pregunta de investigación y reducirlos a códigos. Más bien, en una *primera fase de generación de hipótesis*, se anotan *todos* los significados concebibles para cada segmento del texto, es decir, para cada frase o parte de una frase, según el procedimiento. La generación de hipótesis es estrictamente secuencial, segmento de texto por segmento de texto y a lo largo de una falsación estricta, de modo que sólo queda una estructura integrada de significado, que se ha creado sobre el texto y su estructura secuencial natural y se ha comprobado críticamente sobre el texto. Sólo cuando uno está convencido de que ha anotado todas las hipótesis significativas para la pregunta de investigación en los segmentos cribados y que estas hipótesis se han reducido qua falsificación a una única hipótesis central (preliminar) de estructura de caso, procede a la *falsificación* dirigida en otros segmentos del texto. A continuación, se utiliza el resto del texto, es decir, todos los pasajes del texto no examinados hasta el momento, para la falsación selectiva de la hipótesis estructural preliminar formulada. El objetivo es la falsación de la hipótesis estructural preliminar (de caso). Una buena hipótesis de estructura (de caso) se comprueba especialmente en los pasajes críticos del texto, de modo que se demuestra como una estructura de caso probada y no refutada y constituye el resultado del análisis. Si se produce una falsación, en sentido estricto, hay que volver a la primera fase de generación de hipótesis. Ahí es donde hay que encontrar el error. Todos los pasos descritos se han de volver a recorrer. La hipótesis de

estructura de caso finalmente aceptada describe la estructura de significado latente de la mayoría de las acciones (sociales) contenidas en el texto.

826

Capítulo 9

El Paradigma de Codificación para Analizar Datos Cualitativos

> "La convicción es la creencia de que se está en algún punto de conocimiento en posesión de la verdad incondicional".

Menschliches, Allzumenschliches [630]
Friedrich Nietzsche, 1844–1900

9.1 Codificación cualitativa y análisis de contenido

Dependiendo de la fuente de datos, los datos textuales pueden contarnos los mismos hechos, o incluso los mismos hechos en diferentes formulaciones. Como hemos señalado anteriormente, estos datos son distintas formas de protocolos de sucesos o acciones. Dependiendo de su perspectiva, los registradores representarán un acontecimiento de forma diferente. Esto es lo que hace que los métodos cualitativos sean tan valiosos, porque se puede acceder a perspectivas subjetivas. Sin embargo, dependiendo de las posibilidades personales de expresar las propias experiencias, los protocolos escritos, acústicos, visuales o combinados pueden ser idénticos en significado, pero diferentes en redacción o presentación. A partir de la multitud de posibles combinaciones de palabras de prácticamente cualquier idioma y con el trasfondo de las influencias y experiencias individuales, familiares-biográficas y socioculturales nos sorprendería mucho encontrar los mismos niveles de significado expresados siempre de la misma manera. Por eso es importante extraer de los textos los niveles de significado comparables y compararlos entre sí.

Para ilustrarlo, he aquí algunos ejemplos de pasajes de las transcripciones de entrevistas biográficas con jóvenes, que se recogieron en el marco de un proyecto de la DFG (Huber & Kenntner, 1986):

- "En realidad, el impulso vino de mi madre".
- "Si hubiera habido mucha presión por parte de mi madre, entonces ..."
- "Al principio mi madre no estaba muy entusiasmada; decía, ahora hazlo".
- "Bueno, entonces dejé que mi madre me influyera".
- "Sí, mi madre esperaba mucho de mí".
- "Bueno, debió ser un poco desagradable desde el punto de vista de mi madre".
- "Ya tienes 18 años, me dijo mi madre, ahora puedes venir a casa cuando quieras".
- "Mi madre también fue una persona clave para mí, mi gran modelo a seguir".

Lo que tienen en común estas afirmaciones redactadas de forma diferente es que se refieren a la influencia de la madre en el entrevistado. Por tanto, se podrían abreviar todas estas afirmaciones y representarlas mediante la formulación "*influencia de la madre*", a lo que habría que añadir la calidad de la influencia (por ejemplo, positiva, negativa, neutra) en función de la pregunta de investigación. Como en todos los enfoques de investigación empírica, seguimos un camino de reducción de detalles concretos a niveles

superiores de abstracción y generalización. Lo esencial se hace visible cuando eliminamos "... todos los elementos accidentales del mundo real" (Galtung, 1990, p.98). Sin embargo, el procedimiento metodológico también debe mantener abierto el camino de vuelta a los detalles concretos, coloridos y diversos.

Así, en un proceso de interpretación, buscamos unidades de significado en los datos cualitativos y reducimos cada una a una categoría o código – en el ejemplo anterior, se trata de la *influencia de la madre*. Lo que buscamos depende, por supuesto, de la pregunta de investigación. En el transcurso de este proceso, al comparar las codificaciones dentro de un texto y entre los textos del estudio, se desarrolla un *sistema coherente y consistente de categorías*. Esto constituye la base de los análisis posteriores, en los que buscamos, por ejemplo, conexiones de significado entre los textos, patrones específicos de experiencia de las personas entrevistadas u observadas, similitudes y diferencias entre determinados grupos de personas, etc., para finalmente poder responder a la pregunta de investigación. Se trata de un proceso no trivial que no puede resolverse simplemente de forma algorítmica, sino que requiere muchas repeticiones bajo auspicios cambiantes y un alto grado de creatividad.

Lo descrito aquí corresponde aproximadamente al procedimiento de análisis de contenido latente: *Los textos múltiples se reducen interpretativamente a unidades de significado o unidades de contenido latente y se representan mediante códigos*. En el caso del análisis del contenido manifiesto, es decir, la determinación de la frecuencia de determinadas palabras o secuencias de palabras en el texto, se debe realizar de antemano el trabajo interpretativo, ya que hay que determinar qué palabras se van a contar. Dado que estas palabras representan "algo", el contenido latente también entra por la puerta de atrás. Las palabras clave deben ser palabras que representen un significado específico, es decir, que proporcionen reglas de traducción para determinados aspectos del significado en el texto. Requisito previo para el análisis del contenido manifiesto es disponer de una lista de palabras clave relevantes para la pregunta de investigación y de presuposiciones teóricas o empíricamente probadas de las que se puedan derivar las reglas de traducción. Esto corresponde al proceso cuantitativo de, por ejemplo, *entrenar* un sistema de clasificación en un conjunto de datos para luego *aplicar* este modelo a otro conjunto de datos. Como ejemplo, nos referimos al análisis de las características de los pasajeros del barco "Titanic" respecto a la determinación de las probabilidades de supervivencia (véase el capítulo 5.5.4). Al igual que los códigos tras la interpretación en busca de un contenido latente del texto, las palabras clave para el análisis del contenido manifiesto son representantes del significado. En ambos casos, por tanto, se trata de representantes o códigos en el análisis.

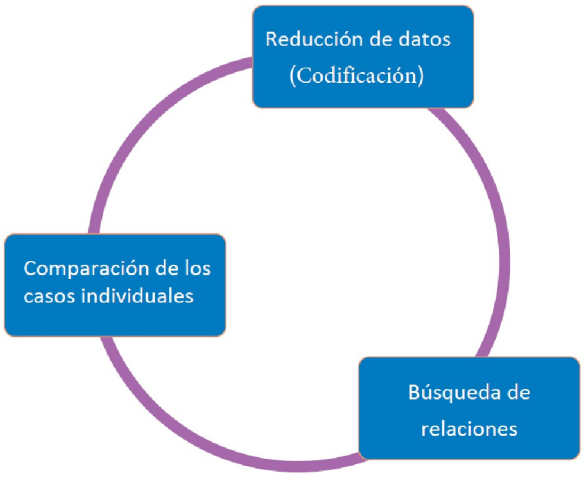


Figura 9.1 *Análisis de contenidos latentes*

Las explicaciones anteriores sobre el análisis del contenido manifiesto y latente ya muestran que se crea una gran cantidad de datos a la vista de los códigos generados, las palabras clave seleccionadas, etc. Para no perder la pista, necesitamos una visión general de los datos. Para no perder el rastro de los datos, debe utilizarse urgentemente el software QDA (s. cap. 11). (véase el capítulo 11.12), que puede gestionar los datos

(texto, códigos, códigos abstractos, búsqueda de secuencias de códigos, etc.).En definitiva, todos ellos funcionan según el mismo principio: permiten gestionar los datos y los códigos para documentar los distintos pasos analíticos y tenerlos siempre a punto. El trabajo interpretativo propiamente dicho está prohibido, porque eso significaría interpretar los textos algorítmicamente y en concreto de forma adecuada al caso en relación con la pregunta de investigación. En la actualidad, ningún programa informático puede (¿todavía?) hacer frente a esta complejidad, aunque los avances ya son asombrosos cuando entran en juego sistemas de interpretación predefinidos y algoritmos de IA, como puede verse, por ejemplo, en la manipulación de material fotográfico, de vídeo y de audio (palabra clave: deepfakes).

9.2 El análisis del contenido latente – Análisis cualitativo de textos

El análisis cualitativo de textos suele dividirse en tres grandes secciones: reducción, reconstrucción y comparación de casos (véase la figura 9.1):

1. En primer lugar se reducen los datos originales a codificaciones como representantes de aquellas unidades de significado que parecen relevantes para la pregunta de investigación.
2. En la fase de conclusión, se utilizan estos códigos para intentar reconstruir las relaciones de significado dentro de los casos individuales del estudio.
3. Comparando los casos en busca de similitudes o diferencias (véase también el capítulo 9.5), se pueden elaborar conclusiones generalizadoras a partir de los datos disponibles.

En los estudios psicológicos, en particular, es interesante la conclusión inductiva de los datos de una persona individual a patrones típicos de su experiencia y comportamiento como conclusiones generalizadoras. Además, a menudo interesa como conclusión si se pueden encontrar puntos en común entre varias personas o situaciones de observación. Un examen detallado de las tres tres fases es el siguiente:

9.2.1 Reducción – Categorización y Codificación

El análisis textual cualitativo comienza con la *reducción de la base de datos* y, al mismo tiempo, de la diversidad de expresiónes que contienen, es decir, las formulaciones lingüísticas de los textos y, además los rasgos paralingüísticos en las grabaciones de audio y las expresiones no verbales en las grabaciones de vídeo. Para ello, se utilizan segmentos de datos más o menos extensos (segmentos de texto, cinta o vídeo o fragmentos de imágenes). A cada uno de ellos se le atribuye un significado negable. A continuación, se adjunta un código al segmento de datos como abreviatura o designación del significado. En lo sucesivo, estos códigos se tratarán como representantes de los segmentos de datos o de las *unidades de significado* de los ficheros. Básicamente se trata de un proceso de categorización, en el que las categorías exactas sólo suelen desarrollarse durante la interpretación de los datos. Sin embargo, se pueden las categorías tomar de un sistema de categorías ya existente. Esto depende totalmente de la pregunta de investigación. En función de las necesidades, se pueden juntar varios códigos en códigos más abstractos o "metacódigos" (Huber & Gürtler, 2012, capítulo 7.3); para cada uno de los cuales se requiere una regla sustantiva justificada en términos de contenido.

9.2.2 Reconstrucción de sistemas de significado subjetivos

A partir de estas unidades de significado, se reconstruyen los sistemas subjetivos de significado de los productores de datos, es decir, los entrevistados, diaristas, observadores, etc. En la *reconstrucción* se buscan vínculos regulares entre las unidades de significado dentro de los archivos de datos que son características de los productores de datos y/o su situación. Esto significa yendo y viniendo entre los datos reducidos (códigos y metacódigos) y el material original. Por un lado, esto refuerza y verifica la coherencia del sistema de categorías. Por otro lado, esto conduce a una comprensión más profunda del caso mediante la identificación de áreas del caso (por ejemplo, declaraciones, reacciones, falta de reacciones, etc.), que se comparan sistemáticamente entre sí.

9.2.3 Comparación de casos

Por último, en una tercera fase, al comparar las percepciones de los sistemas de significado individuales obtenidas de este modo, se identifican invariantes o correlaciones generales en todos o en (sub)grupos individuales de casos individuales (Ragin, 1987). En este caso, se utilizan métodos lógicos de análisis de datos (véase el capítulo 12), que pueden contribuir y simplificar procedimientos complicados mediante el uso de ordenadores.

Es importante señalar que estas fases no son estrictamente delimitables y se suceden linealmente sino que suelen solaparse y desarrollarse en ciclos (Shelly y Sibert, 1992). Un ejemplo – ya durante la reducción del informe de un profesor sobre sus alumnos, se puede pensar en las teorías implícitas de los alumnos sobre este profesor mientras se lee. Así, después de revisar unos cuantos expedientes, se empieza a comparar constantemente los casos entre sí. En el proceso surge claramente un aspecto en el caso de la persona B que se había pasado por alto en el caso de la persona A. Así que volvemos a pasar por la fase de reducción, al menos parcialmente, y así sucesivamente. Luego pasamos a la persona C y utilizamos este trabajo preliminar. Es importante trabajar falsificamente y no sólo limitarse a intentar confirmar los propios supuestos.

En cada una de estas fases, siempre es importante comprobar deductivamente la validez de las categorías desarrolladas inductivamente (véase la Fig. 9.2). Para ello, se utilizan las categorías y las definiciones elaboradas para ellas hasta el momento para recurrir a los datos que ya se han examinado y se intenta aplicar de nuevo la regla de definición actual. Al menos al principio del trabajo de análisis, esto puede ir seguido de revisiones más o menos extensas del sistema de categorías en desarrollo. La ventaja de un sistema de categorías de alcance limitado es que aún puede gestionarse y sigue siendo viable. Esto no excluye que se sigan conservando códigos individuales en el procedimiento basado en software. Pero sin una agrupación hábil y orientada al contenido de los códigos, no se podrá reconocer lo esencial.

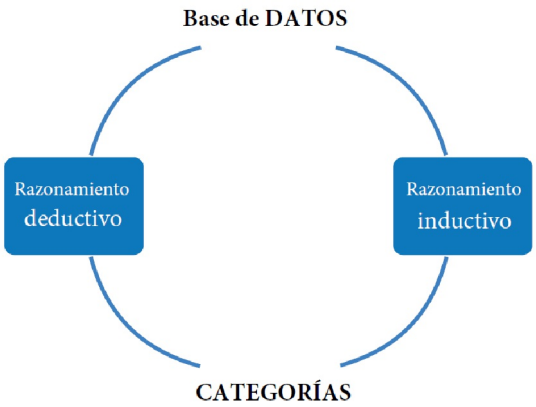


Figura 9.2. *Razonamiento inductivo y deductivo al codificar*

9.2.4 Codificación cualitativa y análisis de contenido

Ya hemos señalado anteriormente que los códigos sirven como representantes del significado de los segmentos de datos, es decir, las *unidades de significado* descubiertas en los archivos. Pero, ¿cómo se encuentran las unidades de significado? En primer lugar, hay que volver a decirlo: Todo depende de la pregunta de investigación y de los datos recogidos para responderla. En el resumen (véase la Fig. 8.1) de los procedimientos de análisis de datos cualitativos, distinguimos tres procedimientos en la búsqueda de contenido latente: *abierto, guiado por hipótesis y cerrado*. O dicho de otro modo: ¿buscamos de forma abierta, tenemos una modesta corazonada o queremos encontrar algo concreto de forma dirigida?

9.2.4.1 Enfoque abierto

Se puede acercarse a los datos completamente abierto, es decir, abierto a cualquier significado que pueda surgir en ellos. Supongamos que un grupo de investigadores se interesa por cómo vivieron los estudiantes de primer curso de una asignatura concreta la transición de la escuela a la universidad tras acabar el bachillerato. Unas semanas después de comenzar sus estudios, los investigadores solicitan una entrevista a una muestra de estudiantes mediante una entrevista sobre la pregunta inicial no específica: "¿Puedes describir cómo experimentas la transición de la escuela a la universidad?" o "¿Qué ha cambiado para ti después de la escuela hasta ahora?". Dependiendo del curso de la entrevista, pueden abordarse preguntas generales adicionales sobre las razones para elegir el campo de estudio y las experiencias, tanto positivas como negativas, útiles o simplemente subjetivamente importantes. A continuación pueden formularse preguntas sobre qué aconsejarían los entrevistados a futuros estudiantes de primer año, etc. De este modo, los entrevistados responden a preguntas abiertas y no canalicen ya sus expresiones en la dirección de las presuposiciones o categorías de los entrevistadores.

En consecuencia, los investigadores tienen que permanecer abiertos en su búsqueda de unidades de significado para poder generar hipótesis sobre su pregunta de investigación a partir de los datos. Esto incluye notar y responder a las desviaciones de las propias expectativas como investigador durante la entrevista. Este enfoque es típico del enfoque de la *Grounded Theory* (teoría fundamentada; Glaser & Strauss, 1967; Strauss & Corbin, 1990; Charmaz, 2012). En el capítulo 9.3 sobre la teoría fundamentada nos referiremos a las especificidades de la codificación inicialmente abierta, después centrarnos gradualmente en el significado central de este enfoque.

9.2.4.2 Orientación por la hipótesis

En muchos casos, los investigadores buscan datos sobre hipótesis muy concretas desde el principio, por lo que utilizan entrevistas semiestructuradas o guiadas. Sus preguntas son básicamente "abiertas", de modo que los entrevistados no pueden responder simplemente "sí" o "no" o "sólo dos veces hasta ahora" o similar. Pero los investigadores aumentan la comparabilidad de los resultados limitando la gama de contenidos de las respuestas en las preguntas y proporcionando así una cuadrícula.

En un estudio sobre las cualidades de liderazgo de los directores de centros escolares (Gento, Huber, Gonzáles & Orden, 2015b, véase también el capítulo 5.5.5), se introdujo a los entrevistados en el tema con una pregunta general, pero después los entrevistadores formularon preguntas abiertas según ocho dimensiones diferentes del liderazgo, especificadas en una directriz. No tendría sentido hablar de un enfoque de teoría fundamentada en el análisis. Al fin y al cabo, más adelante se buscarán en los datos descripciones y evaluaciones específicas de las dimensiones hipotéticas. Es posible que descubramos que los encuestados perciben el tema de una forma ligeramente distinta. Por lo tanto, un análisis cualitativo basado en hipótesis debe, en general, permanecer siempre abierto a cualquier cosa que quede fuera de la cuadrícula conceptual dada en las respuestas. Por lo tanto, un análisis cualitativo guiado por hipótesis debe estar siempre abierto a cualquier cosa que se salga de la rejilla conceptual dada en las respuestas y tenerla en cuenta y, si es necesario, incluirla.

Por lo tanto, es aconsejable definir categorías concretas para cada una de las hipótesis ya al planificar la recogida de datos, por ejemplo al definir las preguntas orientativas para la entrevista, así como antes del análisis del texto (parte deductiva). Por otra parte, durante el análisis del texto es seguro que se descubrirán pasajes en las entrevistas que no pueden asignarse a las categorías predefinidas (parte inductiva). Estas unidades de significado se marcan y las categorías se desarrollan inductivamente a partir del material de datos, vinculándolas a las hipótesis específicas del tema. Por supuesto, en este procedimiento no se pueden descartar modificaciones sustanciales del marco de orientación original. El grado de libertad del investigador y las exigencias a la propia actuación interpretativa aumentan con esta estrategia, al igual que las posibilidades de hacer justicia a las perspectivas específicas de los sujetos de investigación en el análisis.

Tabla 9.1: *Categorías para el análisis de interacción según Flanders (1970)*

Foco	Categoría
El profesor ...	enseña
	da instrucciones
	crítica algo / justifica su autoridad
	acepta emociones
	alaba o alenta
	acepta o utiliza ideas de los alumnos
	pregunta
El alumno ...	respuesta
	inicia un discurso
Algo más	atribución no posible

9.2.4.3 Enfoque cerrado

La forma más sencilla de determinar unidades de significado es utilizar un sistema de categorías ya disponible, sus definiciones y ejemplos de uso a la hora de codificar. Tales sistemas de categorías pueden estar disponibles en el caso de estudios de replicación de la propia investigación previa, o se pueden utilizar sistemas de categorías encontrados en la literatura empírica o en análisis teóricos del área de contenido de interés. Un ejemplo famoso de los años 70 son las categorías para el análisis de la interacción en el aula de Flanders (1970, "Flanders Interaction Analysis Categories", FIAC), que comprenden diez categorías según las cuales se pueden codificar las interacciones entre profesores y alumnos en la observación de clases (véase la tabla 9.1).

Por supuesto, los usuarios tendrán que familiarizarse a fondo con estas definiciones y los ejemplos en situaciones de formación, por ejemplo, utilizando grabaciones de vídeo. En la situación real, cada comportamiento de profesores y alumnos se asigna simplemente a una de las categorías y se registra para su posterior análisis. En este enfoque, no se quiere inferir significados de los datos existentes, sino que sólo se decide qué secciones de los datos corresponden a qué ejemplos de definición de las categorías del sistema. Así pues, la apertura es mínima o inexistente. Si algo no encaja en la rejilla de búsqueda existente, es casi seguro que se pasará por alto. La versión radical de este enfoque selectivo es el uso del cuestionario, posiblemente con preguntas que sólo se pueden marcar ("Multiple choice"; opción múltiple) y que se contabilizan primero en la evaluación.

A la hora de determinar las unidades de significado, siempre hay que tener en cuenta que *estas unidades analíticas deben desarrollarse durante la interpretación de los datos*. Esto también se aplica de forma simplificada a las sistemas de categorías ya disponibles, pero especialmente cuando primero hay que desarrollar categorías interpretativas de forma inductiva para comprender o explicar las experiencias y

acciones de las personas desde su punto de vista subjetivo. Si se intenta acceder a la visión subjetiva del mundo de los interlocutores o los observados, entonces la malla interpretativa del investigador no debe determinar por sí sola qué aspectos de las declaraciones o acciones se captan, posiblemente desvinculados del contexto subjetivo de significado. Sin embargo, tampoco tiene mucho sentido adoptar categorías sin cuestionar los puntos de vista subjetivos de las personas estudiadas cuando se trata de la orientación efectiva de la acción de las *teorías subjetivas* (Groeben, 1986).

La apertura descrita se aplica en particular a los análisis de textos según el enfoque de la *teoría fundamentada* (véase el capítulo 9.3), la formación de teoría empíricamente asegurada o anclada en el objeto según Glaser y Strauss (1967). Sin embargo, este enfoque requiere una cantidad de trabajo considerable, en cuanto hay que analizar más de un conjunto de datos. Puesto que se desea que los resultados individuales sean comparables en una fase más avanzada del análisis, hay que examinar – y esto normalmente varias veces – la coherencia de las unidades de significado y sus codificaciones en todos los archivos y, posiblemente, modificarlas. Aquí es donde el software QDA resulta indispensable para mantener una visión de conjunto.

Un factor de complicación adicional es que el trabajo cualitativo no permite hacer afirmaciones sobre el tamaño necesario de la muestra, como se practica en el marco de la teoría estadística de Neyman-Pearson (véase el capítulo 4.3.3). Hildenbrand (1996; 1999) resuelve este problema "casando", por así decirlo, el análisis de casos individuales con el análisis de secuencias. Basándose en los análisis que ya han tenido lugar, se selecciona el siguiente caso, lejos de arbitrariamente o incluso al azar, sino según el *método del "contraste máximo"*, que como principio es mucho más antiguo que la teoría fundamentada, por ejemplo, se conoce este método del trabajo con colores o, en un sentido más amplio, pertenece a la teoría de la armonía. Esto nos permite encontrar y reconstruir más rápidamente aspectos de campo desconocidos. Así pues, la selección tiene lugar en el nivel de la selección de casos y no tanto en el nivel del análisis de datos. Según Hildenbrand (2006a), con un análisis y una selección de casos cuidadosos, bastan entre 8 y 12 casos para investigar exhaustivamente un campo. Para el nivel de análisis de datos, Miles y Huberman (1994) han vuelto a proponer un compromiso, según el cual, *antes de leer* los textos o revisar los archivos, se establece un marco de orientación muy general, *no específico del contenido*, para la búsqueda de unidades de significado, dentro de luego se decide sobre las unidades específicas según las condiciones del caso individual (ibid., p.18).

> "La generación de una teoría se basa en unos pocos constructos generales que incluyen toda una montaña de detalles específicos. Categorías como 'clima social', 'escena cultural' y 'conflicto de roles', por ejemplo, son etiquetas que pegamos en 'cajones' intelectuales que contienen muchos acontecimientos y comportamientos diferentes. Todo investigador, por muy inductivo que sea su enfoque, sabe qué casilleros pueden entrar en juego en su investigación y qué es probable que haya en ellos. Los casilleros proceden de la teoría y la experiencia y (a menudo) de los objetivos generales que persigue el estudio. La unificación de los casilleros, nombrarlos y tener más claras sus conexiones nos lleva a un marco conceptual de referencia".

En primer lugar, es importante señalar que, incluso cuando trabajamos de forma *inductiva*, nunca estamos completamente abiertos, sino que siempre percibimos el mundo y, por tanto, todos los datos a través de nuestras propias gafas de colores. Esto no puede evitarse. Este compromiso parece justificable en la medida en que la reducción de datos *comienza al principio* de una investigación, incluso en la formulación de la pregunta de investigación. La decisión por un determinado enfoque de codificación, por ejemplo la "*voice listening*" (escucha de voz; Gilligan, Spencer, Weinberg & Bertsch, 2003) o la *codificación de emociones* (Saldaña, 2009), determina el marco de orientación conceptual de los análisis posteriores. No obstante, se debe ser y seguir siendo consciente de que cualquier orientación puede excluir e impedir otros puntos de vista, de modo que los aspectos de los datos que no encajen en el marco de orientación no se consideren más adelante en el análisis. La decisión a favor de algo es siempre también la decisión en contra de todo lo demás (véase el capítulo 11.3). Por lo tanto, los investigadores cualitativos deben permanecer lo más cerca posible de los datos, independientemente de la forma en que quieran determinar las unidades de significado, y permanecer abiertos a todo lo que pueda surgir en los datos en términos de significado. Además se ofrecen asistencia el uso de ordenadores (por ejemplo, para pasajes equivalentes en los textos), trabajar

y debatir en grupo para reducir la propia ceguera profesional, así como llevar diarios de investigación para experiencia primaria como investigador.

9.3 La teoría fundamentada (Grounded Theory)

La apertura al contenido latente de los datos, el significado que puede surgir del espacio "entre líneas" del texto, es el requisito básico de todos los enfoques analíticos que, en la tradición de la Teoría Fundamentada (Glaser y Strauss, 1967), quieren desarrollar teorías sobre los contextos registrados de experiencia y/o acción a partir del material de datos. Los investigadores deben prestar atención a las particularidades del caso con la mayor imparcialidad posible para extraer sus conclusiones y condensarlas en reglas generales. Las regularidades así construidas por inducción o abducción constituyen la base de conclusiones equivalentes en el caso de especificidades que parezcan similares. Sin embargo, el proceso de generación de teorías requiere un constante control deductivo, es decir, comprobar si la regularidad encontrada es igualmente válida o sigue siéndolo en ejemplos de datos similares (véase la Fig. 9.3). De este modo se cierra el círculo, de modo que en la práctica los distintos tipos de inferencia simplemente coinciden y sólo pueden separarse con fines analíticos.

Por ello, Glaser y Strauss presentan la *comparación permanente de casos* como el método central de su enfoque. Como subrayan Shelly y Sibert (1992), los procesos comparativos permanentes de la construcción teórica basada en el objeto *proceden cíclicamente*. Siguiendo su clasificación, se pueden describir estos procesos comparativos y los objetivos subyacentes en el progreso del análisis como se resume en la figura 9.3.

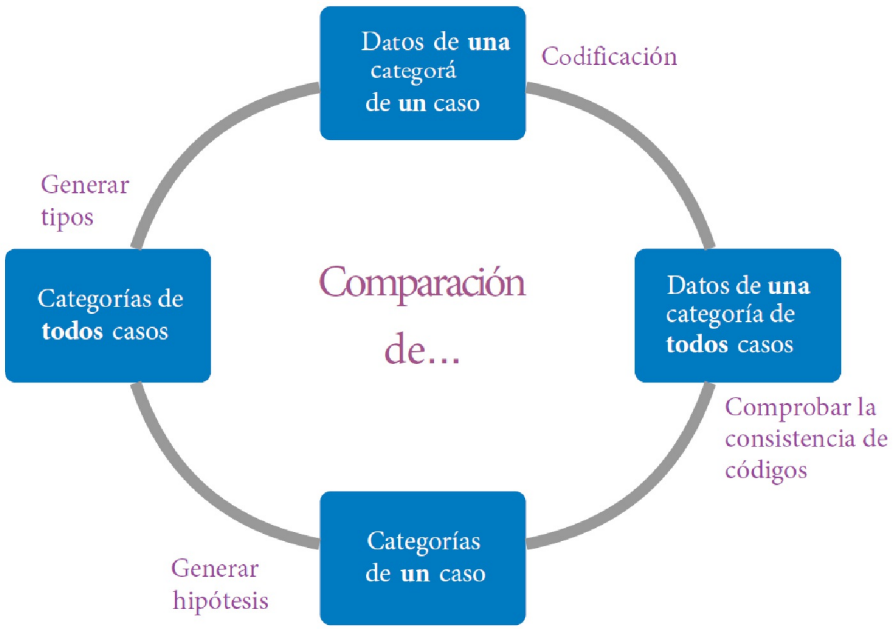


Figura 9.3. *Comparación permanente en el análisis cualitativo*

Tras la decisión básica sobre cómo proceder con el análisis, se concretan las consideraciones y se empieza a codificar una muestra de los textos disponibles. La selección de la muestra puede basarse en diversas consideraciones: similitud, disimilitud, cuestiones específicas de contenido, etc. En concreto, primero se busca dentro de un mismo archivo segmentos de texto, secuencias de sonido o vídeo, fragmentos de imágenes, etc. a los que se puedan atribuir significados específicos adaptados a la pregunta de investigación.

Al hacerlo, se pasa necesariamente de comparar e integrar datos individuales en el conjunto de datos único a comparar las categorías creadas en el proceso en otros conjuntos de datos de la muestra. Así pues, se mueve en pasos analíticos más pequeños y menos omnipresentes dentro del proceso de investigación cíclica para garantizar primero la coherencia de las codificaciones. El objetivo es desarrollar un sistema coherente de categorías que sea a la vez específico para cada caso y lo suficientemente generalizado como para abarcar todos los casos. Es aconsejable anotar y explicitar estas reglas de codificación, a menudo implícitas, incluyendo ejemplos textuales, para al menos algunos códigos: cuándo codificar, cuándo no codificar, etc. No se trata de un esfuerzo superfluo, ya que precisamente esos ejemplos pertenecen a la elaboración o al apéndice de un trabajo de investigación.

En teoría, habría que formular una regla de clasificación clara para cada categoría y poner un ejemplo (positivo) y un ejemplo negativo. Para ir sobre seguro en la investigación diaria, puede adjuntar una nota al texto para cada nueva categoría (si lo admite el software: escriba una nota y guárdela) donde se originó. Esto ayuda a recordarlo más tarde.

Las reglas de codificación así desarrolladas se aplican después a todos los archivos disponibles. En el proceso se topará regularmente con contradicciones y excepciones que incitan a modificar las reglas de codificación. En algunos casos, estas percepciones pueden hacer necesario abandonar el ciclo de reducción/interpretación de datos y volver a entrar en la fase de recopilación de datos en otro movimiento cíclico con el fin de aclarar las cuestiones abiertas de forma específica.

Si el uso coherente de los códigos parece estar garantizado, es posible buscar específicamente conexiones sistemáticas en los códigos y entre ellos, conexiones que probablemente ya se hayan advertido durante la codificación de textos individuales o el examen de la coherencia de la codificación. Una vez más, si se observan tales conexiones, deben "graparse" directamente en el pasaje de texto correspondiente, ya sea en un diario de investigación y/o con software QDA como una nota, es decir, un *memo* (Huber & Gürtler, 2012, pp.143-148). Así que ahora se empieza a formar hipótesis sobre las conexiones en los archivos individuales, como "Si la educadora E observa el evento A o B, entonces aplica la opción de acción X". O "Cuando el profesor Z habla de la falta de motivación de sus alumnos, menciona la influencia de los medios digitales". Esto a su vez requiere el cambio sistemático y cíclico entre los niveles de códigos y datos originales (texto, vídeo, audio, imagen). Se pueden utilizar las experiencias en esta fase, por ejemplo las contradicciones en las interpretaciones, también como ocasión para un bucle analítico de vuelta a través de las fases de codificación y garantía de coherencia. También puede resultar necesaria una nueva recopilación de datos con preguntas moduladas, situaciones de observación, etc. Sea como fuere, se requiere perseverancia y tenacidad para no aceptar precipitadamente una interpretación que puede no ser válida.

Suponiendo que se haya registrado un número suficiente de casos en el estudio, se puede concluir con un intento teóricamente justificado de agrupar los casos individuales en *tipos* sobre la base de la comparación y la integración de las categorías seleccionadas. Los tipos (Kelle & Kluge, 1999) son configuraciones de códigos casi complejas, que en sí mismas y a lo largo de su interpretación son significativas y favorecen el conocimiento de la pregunta de investigación. Esto les confiere un buen efecto de memoria y, en consecuencia, facilita su comunicación. Como resultado, obtenemos entonces una diferenciación de los casos según conjuntos de combinaciones de rasgos típicos. Por ejemplo, en un estudio en el que se intentara registrar las explicaciones de los profesores sobre la falta de motivación de sus alumnos para aprender, se podría pedir a los profesores que dieran unas explicaciones. Después se podría clasificar a los profesores según los *tipos de explicación* y, como consecuencia práctica, se podrían diseñar intervenciones pedagógicas adecuadas o medidas de formación adicional en un paso posterior.

Especialmente para los principiantes en la interpretación de datos cualitativos anclados en la materia (= enfoque de grounded theory), puede ser un problema conciliar la *exigencia básica de apertura* a todas las unidades de significado que aparecen en el material de datos con la necesidad pragmáticamente justificada de elaborar los significados centrales relevantes para la pregunta de investigación en *un número manejable de categorías significativas* – en resumen: mantener la visión de conjunto. Si alguien – como se ha observado en un caso real – ha utilizado más de 1.500 códigos en la interpretación de 20 entrevistas de aproximadamente una hora, literalmente "dejará de ver el bosque por los árboles" y, por lo tanto, tendrá grandes dificultades para reconstruir los sistemas de significado de los textos de forma comparable. En este sentido, merece la pena echar un vistazo rápido al número de categorías creadas de vez en cuando. Asimismo, los

analistas pueden preguntarse si tienen que crear constantemente nuevas categorías para cada nuevo caso (por ejemplo, entrevista, ...) o si no sería mejor utilizar las existentes a partir de material ya codificado o modificarlas discretamente.

Por ello, los representantes del enfoque de análisis de la teoría fundamentada (Strauss & Corbin, 1990; Charmaz, 2006) sugieren distinguir distintas fases en el proceso de interpretación. Partiendo de (1) la *codificación abierta o inicial*, se pasa a (2) la *reducción de los códigos* según distintos criterios (por ejemplo, codificación *focalizada* frente a *axial*); y se termina (3) con la elaboración e integración con respecto a una *categoría central abstracta*. En detalle, esto tiene el siguiente aspecto:

9.3.1 Codificación

En primer lugar, estrechamente orientado al material de datos, se deben reducir las respuestas de los entrevistados a las preguntas o impulsos de los entrevistadores, expresadas en muchas formulaciones diferentes, a conceptos más abstractos, es decir, los códigos. Strauss y Corbin (1990) hablan aquí de "codificación abierta", Charmaz (2006) llama a esta fase "codificación inicial".

El *requisito de permanecer cerca de los datos* durante la codificación resulta ser crítico y todo un reto. Por un lado, esto implica que es esencial trabajar con las transcripciones o grabaciones originales, pero no con resúmenes parafraseados de las entrevistas, por ejemplo, y menos aún con las propias grabaciones durante las entrevistas. Esto se debe a que no sólo es probable que las paráfrasis y las grabaciones contengan interpretaciones espontáneas basadas en conocimientos cotidianos u opiniones preconcebidas y posiblemente incluso prejuicios. Esto ocurre de forma habitual e inconsciente. Antes se parafraseaba ampliamente practicado en el curso del conocido libro de texto sobre análisis de contenido cualitativo de Mayring (1995). Históricamente, esto también tenía sentido, ya que en aquella época no se disponía de ordenadores rápidos para la gestión de datos y códigos, y la paráfrasis inicial podía conducir a una reducción justificada del texto. Hoy en día, este argumento no es válido y cada categoría, independientemente de su nivel de abstracción, se debería poder referir a la fecha original en cualquier momento. Hoy llamaríamos a la codificación de las propias paráfrasis "interpretación de los propios puntos de vista y (pre)supuestos". Pero en realidad queremos interpretar los datos originales y no lo que hemos hecho de ellos. Esto sólo puede hacerse utilizando software de tipo QDA. El análisis posterior de este material parafraseado sólo conduciría a una "interpretación de interpretaciones" y el acceso a la visión subjetiva original de los entrevistados quedaría claramente obstruido y distorsionado, y las conclusiones se extraerían y justificarían de forma inadecuada.

Para los principiantes en el análisis cualitativo, así como para los que se familiarizan con nuevas áreas de contenido, siempre surge la pregunta: "¿Cómo y dónde se definen las unidades de significado en los archivos que se van a reducir mediante códigos?". Weber (1990) se refiere a seis posibilidades en general para definir unidades de análisis, a saber

1. Palabras,
2. significados de palabras
3. frases,
4. temas,
5. secciones y
6. el texto completo (posible, por ejemplo, para titulares, resúmenes, cartas cortas al editor, etc.).

Charmaz (2006) recomienda, entre otras cosas, proceder a una codificación inicial *palabra por palabra*, *línea por línea* o *acontecimiento por acontecimiento*. La finalidad de este procedimiento es evitar pasar nada por alto y nombrar todos los pensamientos, acontecimientos, etc. contenidos en los datos. Es aconsejable buscar conceptos abstractos para los códigos ya en esta primera fase. Por ejemplo, el concepto de *relaciones sociales* puede servir para diversas formulaciones que hablen de la relación con amigos, familiares, otros

estudiantes, etc. Resulta práctico, especialmente para los datos verbales (entrevistas, conversaciones, etc.), especificar códigos en el nivel de contenido añadiendo un "*cómo se dice*" al "*qué se dice*". Al nombrar los códigos, esta categoría superior debe utilizarse justo al principio para facilitar su localización posterior. Para nosotros, por tanto, hemos introducido la distinción entre *categorías "de qué" basadas en el contenido* y *categorías "de cómo" estructurales*. Categorizamos ambos niveles de forma paralela. De este modo, más adelante podremos, con un esfuerzo razonable, categorizar tanto por tema y contenido como por estructura y modo.

En una diferenciación similar de esta forma de idea, se ha mostrado una recomendación de Strauss y Corbin (1990) de buscar diferenciaciones en las categorías utilizadas al *codificar en abierto*, especialmente si son de carácter general. Estos autores aconsejan analizar cada código una vez introducido – independientemente de los datos concretos – para ver qué características incluye y en qué dimensiones se definen dichas características. Entienden por características de una categoría las propiedades o atributos de los fenómenos incluidos en ella. Por dimensión entienden las posibles expresiones o localizaciones de una característica en un determinado continuo. La Tabla 9.2 ilustra tales posibilidades de diferenciación utilizando el ejemplo del código "*relaciones sociales*", que se utilizó en un estudio de entrevistas de Kenntner, Fischer y Huber (1988) sobre autoconceptos de alumnos (véase la Tabla 9.2).

Tabla 9.2: *Dimensiones de la categoría "relaciones sociales"*

Código	Característica	Dimensión
Relaciones sociales	Calidad	bueno ... malo
	Intensidad	alta ... baja
	Duración	corta ... longa
	Frecuencia	siempre ... rara
	Curso	subiendo ... hacia abajo
	Personas	Madre, padre, hermano/as, amigo/as, ...

Concretamente, se puede organizar dicha división para la denominación de códigos en el ordenador de la siguiente manera utilizando abreviaturas, aquí `sB_:`

```
sB_Calidad, sB_Intensidad, sB_Duración, sB_Frecuencia, sB_Curso, sB_Personas
```

De este modo, los nombres de códigos se pueden estructurar con elegancia. Si tomamos el ejemplo de *diferenciar entre códigos de contenido y códigos estructurales*, por ejemplo cuando un profesor interactúa con los alumnos, podrían resultar las siguientes categorías. Aquí el prefijo I_ significa contenido y S_ para estructura. Mientras que con el contenido nos preguntamos *qué* hace el profesor, con la estructura nos preguntamos *cómo* lo hace. Ambas juntas dan una imagen vívida de una interacción social. A continuación examinamos algunos códigos ficticios que corresponden a este esquema:

```
I_responder, I_intervenir, I_enseñar-en-la-mesa, I_hacer-tareas,
I_dirir-trabajo-individual, I_dirir-trabajo-compañero-o-grupo, ...

S_enfadar, S_interesar, S_distrair, S_curioso, S_amistoso,
S_agresivo, S_rechazante, S_neutral, ...
```

Si ahora codificamos dos veces, es decir, tanto el contenido como la estructura, tendremos al mismo tiempo dos perspectivas del material de datos, que luego podremos mezclar a voluntad. Esto es aún más cierto en el caso de pasajes especialmente críticos, es decir, interesantes, del material textual que debemos examinar más detenidamente (véase la exigencia opuesta de igualdad de trato en el análisis secuencial, véase cap. 11.2). Por ejemplo, una acción, como responder a un alumno, puede ocurrir tanto de forma amistosa como negativa. Si sólo pedimos una perspectiva, la pasamos por alto. Si este trabajo de codificación se realiza

inmediatamente en la primera pasada, el material ya está muy bien ordenado antes de que nos atrevamos a reducir y abstraer las categorías generadas.

9.3.2 Reducción

En una segunda fase, se deben reducir los propios códigos de forma que las categorías relevantes para la pregunta de investigación se elaboren a un nivel de abstracción adecuado. Un requisito previo para ello es normalizar o generalizar la formulación de los códigos, que representan la misma categoría de significado. Es mejor hacerlo ya en la primera fase. Por lo tanto, se trata de concentrarse en los códigos que son significativos para el análisis de los datos desde la perspectiva más estrecha de la pregunta de investigación. Charmaz (2006) habla por tanto de "codificación focalizada" y la distingue de la "codificación axial" según Strauss y Corbin (1990), que intenta desarrollar jerarquías de códigos. Las jerarquías de códigos se crean mediante la subordinación de códigos relevantes para la investigación y relacionados con el significado como subcategorías bajo una categoría superior según una regla de asignación.

Sea cual sea, la opción elegida debe quedar claro que aquí entran en juego fuertes suposiciones sobre la realidad: ¿imaginamos los códigos como iguales o como jerárquicos? ¿Cómo vemos la influencia en un sistema de códigos jerárquicos en contraste con códigos nebulosos? Puesto que se deben tomar estas decisiones siempre de un modo u otro, no se trata, como suele ser el caso, de evitarlo, sino de tomar decisiones conscientes y tenerlas en cuenta para conclusiones posteriores y comunicarlas de forma transparente al mundo exterior.

En todos los casos, en esta fase se toman decisiones para las que se necesitan percepciones de los significados y contextos de significado de los textos, como las que podrían obtenerse en la fase de codificación abierta. Por un lado, estas percepciones ayudan a encauzar la interpretación de los datos de acuerdo con la pregunta de investigación, pero, por otro lado, también ayudan a no limitar la interpretación hasta tal punto que no se tengan en cuenta aspectos inesperados y de mayor alcance.

Aquí recomendamos que las categorías abstractas recién creadas reciban un prefijo, por ejemplo un AA_ al principio o un ZZ_ si queremos que los códigos se coloquen en una posición determinada en el registro de códigos. El orden en el registro de códigos de un programa informático suele resolverse alfabéticamente. Alternativamente, podemos introducir MC_ para metacódigos o cualquier otra cosa que tenga sentido y sea práctico para nosotros. Lo único relevante es encontrar un sistema que nos permita trabajar lo más eficientemente posible con nuestros propios códigos. Dado que el ordenador se encarga de la gestión de los datos, no necesitamos eliminar los códigos anteriores. No interfieren. Como ya se ha dicho, sólo debemos elegir y ordenar las designaciones para no perder la visión de conjunto.

Como estrategia para ayudarnos a encontrar vínculos sistemáticos entre los detalles de los datos y sus códigos correspondientes, Strauss y Corbin (1990, p.99) proponen un "modelo paradigmático". Este modelo exige vincular un fenómeno dado en los datos que es central para la pregunta con sus antecedentes (causales), el contexto de su ocurrencia, las condiciones que intervienen, las posibles estrategias de acción o interacción y las consecuencias. La tabla 9.3 muestra el modelo en términos concretos utilizando el ejemplo del código "alegría" (Huber, 1992). Estas listas esquemáticas de posibles conexiones conceptuales pueden utilizarse como heurística para resumir códigos en el sentido de una superordinación o subordinación "axial". Además, un esquema de este tipo puede estimular a buscar aspectos en los datos que de otro modo podrían pasar desapercibidos.

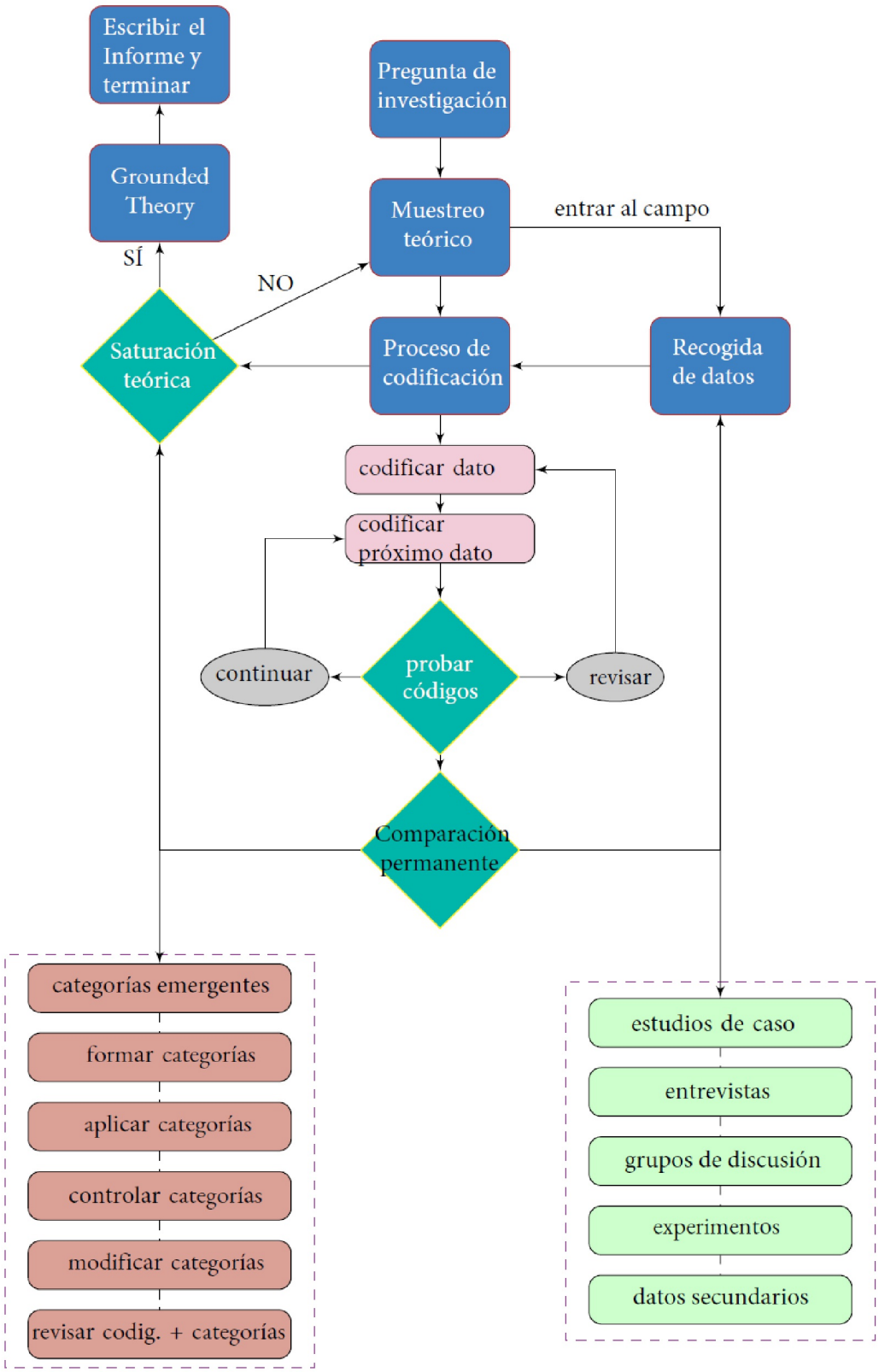


Figura 9.4. *Grounded Theory (Procesos y pasos)*

Tabla 9.3: *Modelo paradigmático según Strauss y Corbin (1990)*

	Código "Alegría"	
Condiciones causales	Encuentro con personas, eventos. objetos	
Fenómeno	Alegría	
		Dimension
Características	Duración	persistente ... a corto plazo
	Intensidad	Intensidad fuerte ... débil
	Referencia social	afirmativa ... negativo
	Referencia futura	fuerte ... débil
Condiciones interviniendas	Experiencias raras (pocos motivos de alegría hasta ahora) frente Experiencias frecuentes	
	Edad de la persona	
Estrategias de acción/ interacción	Buscar contactos sociales	
	Compartir alegría con otros	
	Apertura a los demás	
Consecuencias	Elevación emocional	
	Vivacidad	
	Revalorización del pasado	

9.3.3 Codificación selectiva

Por último, la tercera fase, la "codificación selectiva" (Strauss & Corbin, 1990), representa el verdadero reto del enfoque de la teoría fundamentada. Ahora, en un nivel de abstracción más elevado, se debe elaborar una categoría central general a partir de las interpretaciones (categorías) anteriores. Para ello, estos autores recomiendan cinco pasos (Strauss & Corbin, 1990, p.117s., cursiva en el original):

> "El primer paso consiste en elaborar un concepto para la presentación final, una *story line*. El segundo es vincular las *categorías subordinadas* en torno a la *categoría central* utilizando el *modelo paradigmático*. El tercer paso consiste en relacionar las *categorías entre sí a nivel dimensional*. El cuarto paso consiste en *validar* estas relaciones utilizando los datos proporcionados. La página quinto y último paso consiste en *elaborar* aquellas categorías que necesitan un mayor refinamiento y/o desarrollo.

Como debería haber quedado claro en las observaciones anteriores y en el esquema de la figura 9.2, estas secuencias de pasos son una descripción esquemática del análisis de datos cualitativos. En la realidad del proceso de investigación, siempre es necesario alternar entre la generación abductiva e inductiva de categorías y la verificación deductiva a partir de los datos disponibles. En el caso de la teoría fundamentada, esto se puede adaptar terminológicamente (véase la figura 9.4). Y de vez en cuando hace falta creatividad y valor para hacer algo nuevo. A veces incluso contar ayuda aquí, como hacen Miles y Huberman (1994) en el contexto de sus análisis de tablas.

9.4 Análisis de tablas según Miles y Huberman

Como estrategia más importante para estructurar el contenido textual Miles y Huberman (1994) recomiendan y utilizan en su introducción al análisis de datos cualitativos la forma de ordenar las unidades de sentido o los códigos correspondientes en *tablas* o *matrices*. Esto permite transformar y comprender mejor la configuración *secuencial* de afirmaciones, pensamientos y opiniones en el transcurso del discurso o en la sucesión del texto en una configuración *simultánea* – una tabla – lo que los hace más manejables. Las matrices o tablas ayudan, según los autores (ibíd., p.93s.) a

- obtener o mantener una visión de conjunto, ya que los datos y el análisis se presentan de forma resumida,
- reconocer rápidamente dónde se necesitan análisis adicionales más detallados,
- comparar datos e interpretaciones, y
- comunicar los resultados de la investigación a otras personas y hacerlos comprensibles.

Encontrará más orientaciones sobre la organización de datos verbales en tablas o matrices, junto con numerosos ejemplos, en Miles y Huberman (ibíd., capítulos IV, V y VI). La doble determinación del contenido (segmentos de datos) de las celdas de una tabla de este tipo determina la interpretación de los resultados que la mera constatación de una acumulación de proximidad espacio-temporal de inicialmente de categorías independientes. Por otra parte, la construcción de una matriz de análisis requiere una mayor aportación conceptual, es decir, es necesario un cierto avance en el proceso de análisis para construir una matriz de análisis.

Para concretar, recurrimos a un ejemplo semi-ficticio que hemos recopilado a partir de las entrevistas a profesores de Marcelo (1991) y Zabalza (1991) e incluido en los ejemplos de datos del paquete de software AQUAD 7/8 (Huber, 2019). Para demostrar el análisis de la tabla (véase la Tab. 9.4), comparamos los datos de entrevistas ficticias, y suponemos que éstas se realizaron a 20 profesores de primaria y 20 de secundaria. Con el nivel escolar como característica *singular*, marcamos las columnas de una tabla de representación de los códigos utilizados. Para ello, introducimos los *códigos de perfil* (Huber & Gürtler, 2012, cap. 1.4) `/nivel de primaria` y `/nivel de secundaria` para definir las columnas de una tabla con la que queremos poner a prueba la hipótesis de que existen diferencias entre los profesores de los dos niveles escolares con respecto al ambiente escolar (código conceptual: `ambiente escolar`), los `colegas`, los `problemas` en el aula y las `reflexiones sobre uno mismo`. Los códigos conceptuales `ambiente escolar, colegas, problemas` y `yo mismo` definen las filas de la tabla. Cuando iniciemos el análisis de la tabla, el programa recogerá todos los segmentos del archivo en los que los profesores de primaria hablen sobre el ambiente escolar. En la columna de al lado, leemos las declaraciones de los profesores de secundaria sobre el mismo tema. En la fila siguiente de la tabla, encontramos primero las declaraciones de los profesores de primaria sobre sus colegas, en la columna contigua las declaraciones correspondientes de los profesores de secundaria, y así sucesivamente. Para una primera visión de conjunto, podríamos dejar que el software introdujera en las celdas sólo las frecuencias de los códigos correspondientes, en lugar de los segmentos de texto concretos.

Sin embargo, a más tardar con este intento, uno se da cuenta de que al menos dos de los códigos mencionados no son lo suficientemente sofisticados para la comparación que se pretende. Los códigos conceptuales no específicos, como el ambiente escolar y los compañeros, pueden bastar para una primera diferenciación aproximada de los temas relevantes. Pero limitarse a señalar que los profesores hablan del ambiente escolar y de sus colegas apenas revelará diferencias. Tenemos que calificar estos códigos adicionalmente, es decir, indicar *cómo* hablan los profesores de estos contenidos, por ejemplo, ¿positivo o negativamente? En un paso posterior, podríamos examinar si los profesores, cuando hablan de un ambiente escolar positivo o negativo, ¿hablan también de sí mismos o de otros problemas, etc.? Podemos hacer lo mismo con los comentarios sobre los colegas. De este modo y situada de cómo los códigos conceptuales están vinculados entre sí y en la forma en que se utilizan realmente. Podríamos dar un nombre propio a estas

combinaciones de códigos y obtener estructuras de códigos complejas (véase más adelante), en las que se han incorporado presuposiciones significativas.

Tabla 9.4: *Frecuencias de codificaciones – presentación como tabla*

		Códigos de perfil	
Códigos conceptuales		/Escuela prim.	/Escuela secundaria
	Ambiente escolar (+) positivo	9	12
	Ambiente escolar (–) negativo	1	2
	Colegas (+) positivo	14	11
	Colegas (–) negativo	2	1
	Problemas	8	13
	Si mismo	10	15

El *recuento de la frecuencia de las codificaciones* arroja diferencias llamativas en los dos últimos códigos a primera vista, que podrían aclararse más con las representaciones gráficas de la estadística exploratoria (s. cap. 5). En cualquier caso, hay que volver a mirar las correspondientes declaraciones de la entrevista de cerca y compararlas entre sí. Con ayuda del software (Huber y Gürtler, 2012, cap. 6.4f y cap. 8), estos segmentos de texto pueden encontrarse rápidamente e introducirse en las celdas de la tabla. Es posible que no haya ninguna diferencia sistemática aquí, por ejemplo, debido a una formación diferente de los profesores para las dos ramas escolares. Por ejemplo, podría ser que la muestra con mayor número de profesores contiene un único profesor que experimenta problemas personales específicos en su profesión y, por tanto, habla de ellos con más detalle en la entrevista que sus colegas. Para un análisis más detallado, merece la pena utilizar no sólo los códigos de perfil, sino también los códigos de hablante, con el fin de determinar la distribución de las categorías no sólo entre las dos ramas escolares, sino también entre los individuos. En este paso abordamos de nuevo el análisis cuantitativo y exploramos el AED (s. cap. 5), así como la combinación lógica (s. cap. 12), qué combinaciones de códigos buscamos, por ejemplo, profesor ABC con ambiente escolar (+) positivo, y otras expresiones (ficticias) como abordar contenido emocional (+) o escuchar en lugar de hablar solo, etc. Podríamos agrupar más las frecuencias, representarlas gráficamente, etc. Si tuviéramos hipótesis claras, podríamos incluso hacer estadística inferencial o probarlas específicamente en una nueva muestra. El análisis de tablas según Miles y Huberman es, por tanto, una interfaz entre CUAN y CUAL y nos lleva directamente a los métodos mixtos.

Dichas matrices cumplen los requisitos de estructuración de los datos, sobre todo porque ordenan la amplia información de forma clara y la presentan *simultáneamente*, en lugar de *secuencialmente* como en los archivos originales. Esto permite comparaciones rápidas con los resultados de otras personas, situaciones, momentos, exámenes, etc. Si es cierto que una imagen dice más que 1.000 palabras, una presentación en forma de tabla sustituye al menos a 100 palabras, sobre todo porque puede hacer visibles conjuntamente los datos y el proceso analítico. En la combinación de códigos conceptuales, que se agregan y cuentan mediante complejas reglas de búsqueda en códigos vinvulados que contienen un nivel muy alto de significado junto con reglas de interpretación e hipótesis explicables.

Sin embargo, parece necesaria una advertencia: Incluso una presentación tabular muy detallada de los datos reducidos no puede hacer más que proporcionar una buena base para conclusiones posteriores. En ningún caso, sin embargo, puede sustituir a las conclusiones necesarias y al esfuerzo de análisis asociado. No se interpreta automáticamente en función de las frecuencias: que algo sea frecuente no significa necesariamente que sea importante, y viceversa – porque algo sea poco frecuente deja de ser importante, y viceversa. Una tabla no es una confirmación.

La representación matricial alcanza sus límites cuando no sólo se utilizan características singulares como el nivel escolar o el sexo como criterios de clasificación, sino que también se intenta clasificar según combinaciones específicas de unidades de significado, por ejemplo, subdividiendo las columnas de la categoría de *tipo de escuela* según el *sexo* de los profesores, y quizá también según los grupos de *edad*.

Una matriz bidimensional y, a más tardar, tridimensional se vería desbordada por las exigencias de tales análisis. Una disposición de matrices más diferenciadas o especializadas se volvería rápidamente inmanejable – aparte de que con cada diferenciación de este tipo aumenta el número de entrevistados tendría que aumentar. Al fin y al cabo, todo el mundo debería poder opinar. Sin embargo, la selección de la muestra se aproxima entonces a criterios cuantitativos, que podrían resultar muy obstructivos.

El paso al análisis cuantitativo se define por la cantidad o la cobertura combinatoria de todas las combinaciones de características y no (sólo) a través de la calidad. Sin embargo, hay categorías como el sexo, la persona, el tipo de escuela, etc. como criterios distintivos, siempre que no nos excedamos subdividiéndolos en jerarquías. Este sería el caso, por ejemplo, como se ha descrito anteriormente, si diferenciamos por tipo de escuela según el grupo de edad y, a continuación, por grupo de edad según el sexo, y así sucesivamente. Una matriz multidimensional de este tipo se presta eminentemente a oscurecer lo relevante y a confundir de forma totalmente inapropiada.

Tomemos el ejemplo de la tabla 9.5, donde tendríamos que combinar 4 * 2 * 4 * 9 combinaciones de códigos y ponerlo en una sola tabla. Después de todo, son 288 combinaciones de códigos diferentes, y necesitamos al menos una persona por celda, porque la lógica cuantitativa empieza a entrar en juego aquí, para que no tengamos espacios en blanco. Así que antes de insertar códigos conceptuales y contarlos, nuestra tabla necesita datos para todas las combinaciones de códigos. Tendríamos que realizar al menos 288 entrevistas. No imprimimos lo que esto podría parecer por razones de espacio. Pero un poco de código R da una impresión (`ptIII_qual_code-paradigm_table-analysis.r`):

```
#G = German kind of schools R-Code
kindofschoolG <- c("Hauptschule", "Realschule",
"Gymnasium", "Sonderschule")
sex <- c("f","m")
age <- c("[25-35]", "[35-45]", "[45-55]", "[55-67]")
subject <- c("music", "naturalsciences", "language", "math", "religion",
"sport", "art-and-craft", "IT", "history-and-politics")
tab <- expand.grid(kindofschoolG=kindofschoolG, sex=sex,
age=age, subject=subject)
```

Sólo miramos el principio y el final de la tabla:

```
> head(tab)
kindofschoolG sex age subject
1 Hauptschule f [25-35] music
2 Realschule f [25-35] music
3 Gymnasium f [25-35] music
4 Sonderschule f [25-35] music
5 Hauptschule m [25-35] music
6 Realschule m [25-35] music
> tail(tab)
kindofschoolG sex age subject
283 Gymnasium f [55-67] history-and-politics
284 Sonderschule f [55-67] history-and-politics
285 Hauptschule m [55-67] history-and-politics
286 Realschule m [55-67] history-and-politics
287 Gymnasium m [55-67] history-and-politics
288 Sonderschule m [55-67] history-and-politics
```

Si lo desea, puede aplicar

```
dim(tab)
tab
```

para mostrar la tabla completa, pero aquí vamos a ver *sólo* las categorías, que tendríamos que rellenar *primero* con datos. Tal procedimiento lleva la lógica cualitativa al absurdo.

Tabla 9.5: *Tabla de análisis innecesariamente complicada*

Códigos de diferenciación	sub-categoría	k
Tipo de escuela	Hauptschule, Realschule, Gymnasium, Sonderschule	4
Sexo	feminina, masculino	2
Grupo de edad*	[25–35], [35–45], [45–55], [55–67]	4
Asignatura	Musica, Sciencas naturales, Lengua, Matemática, Religión, Deporte, Artes y oficio, IT, Historia y Politica	9
$\sum$	4 * 2 * 4 * 9 Combinaciones	288

* correspondiente a la experiencia profesional

Volvamos a un enfoque sensato de los datos y las tablas. Es importante que podamos contarlo todo. Esto se aplica no solo a las categorías, los códigos de hablante o de perfil, sino también a categorías más abstractas (metacódigos, Huber & Gürtler, 2012, capítulo 7.3) así como a secuencias y combinaciones de códigos. Sin embargo, para encontrar estructuras necesitamos la identificación de relaciones entre categorías y no meras frecuencias.

9.5 Identificación de estructuras, conexiones y tipologías

Hasta ahora se han esbozado las fases típicas del análisis cualitativo de textos. Sin embargo, no se debe asumir que este análisis es un proceso lineal. La metáfora del *círculo hermenéutico* transmite que la secuencia de la búsqueda de unidades de significado en los textos, a partir de la cual se reconstruyen y se comparan los sistemas subjetivos de significado, no es un proceso lineal, sino es básicamente un proceso interminable. Este proceso avanza sucesivamente en el tiempo y, además, en la adquisición de conocimientos, por lo que precisamente *no* es un círculo, como subraya Stegmüller (1975). En contraste con el ideal del círculo hermenéutico, especialmente popular en los círculos pedagógicos, debemos, sin embargo, mostrar cierta eficacia y orientación hacia los resultados. No queremos avanzar hacia el infinito, sino ver resultados razonables en un tiempo manejable. Esto requiere decidir en un momento dado si se dispone de suficientes datos e interpretaciones para responder adecuadamente a nuestra(s) pregunta(s) de investigación. De lo contrario, por supuesto, siempre podemos meter el dedo al agua y removerlo enérgicamente: algo saldrá. Pero dudamos que esto revele algo *sustancialmente* nuevo.

Una vez que se ha pasado por primera vez por la secuencia de pasos elementales del análisis cualitativo, uno está más familiarizado con los textos o sus productores y sus teorías implícitas o con las influencias que determinan la experiencia y el comportamiento en las situaciones sociales de interés que al principio de la evaluación. Con estos conocimientos más profundos, se puede comenzar de nuevo la interpretación y mejorar todo lo que ahora pueda reconocerse retrospectivamente como un malentendido o una interpretación errónea anterior. Así se crea la coherencia necesaria de los códigos (sistemas de categorías) y las hipótesis que se ponen a prueba en el texto. Al mismo tiempo, podemos trabajar con todas las notas y apuntes que quedan por ahí. A continuación entraremos en detalles sobre la segunda fase en particular, pero sin olvidar que en cada paso hay que volver a asegurar deductivamente la validez de nuestras generalizaciones ya realizadas. Para ello, deducimos datos específicos del paso anterior del análisis y buscamos las pruebas correspondientes o, concretamente, buscamos contrapruebas.

Por supuesto, dependiendo de la pregunta de investigación, uno no se contentará con una lista bien fundamentada de codificaciones como resultado de un análisis cualitativo. Eso correspondería a una enumeración aislada de códigos inconexos entre sí. Casi siempre se buscan *estructuras* en los códigos y patrones y conexiones entre determinados códigos. Por último, a menudo se intenta ordenar el material de datos por *tipos*. La búsqueda de correlaciones es importante porque, de lo contrario, al final sólo se dispone de una lista de frecuencias de los códigos utilizados que, sin embargo, no permite sacar ninguna conclusión sobre las estructuras efectivas. Esto se debería evitar en la medida de lo posible. La práctica demuestra, sin embargo, que sólo unos pocos llevan a cabo realmente estos pasos para crear no sólo un sistema de categorías, sino un sistema de categorías interconectado e interrelacionado que permita poner a prueba incluso hipótesis complejas sobre el texto.

A continuación veremos tres relaciones básicas de codificaciones que deben tenerse en cuenta al buscar estructuras. Todas las variantes de esta búsqueda tienen en común que no sólo hay que tener en cuenta una categoría o los segmentos del texto representados por ella, sino dos o más categorías y *la relación definida entre ellas*, para lo que no existe un criterio natural. Con la *codificación guiada por la teoría*, es decir, *basada en hipótesis* (véase más arriba, Huber y Gürtler, 2012, capítulo 6.4.1), las hipótesis que establecen el marco para el desarrollo de categorías proporcionan indicaciones de ciertas relaciones entre las categorías. En la *codificación abierta*, elaboramos estas hipótesis a lo largo del proceso de codificación. Como memos, registramos los orígenes y las ideas sobre su aparición directamente en el texto. En definitiva, se trata de lo mismo: las hipótesis, operacionalizadas como secuencias de códigos, se buscan y se encuentran en el texto – o no. En ambos casos se trata de información equivalente. Como ocurre a menudo, no se trata de confirmar las propias presuposiciones, sino de una seria reconstrucción de patrones.

Las estrategias generales de búsqueda de estructuras de significado en el material de datos prestan atención a las tres características siguientes: *Jerarquías, Secuencias* y *Agrupaciones*.

9.5.1 Jerarquías de categorías individuales

Tomemos como ejemplo de tal concepto marco un estudio del desarrollo en la adolescencia, que está orientado hacia el modelo de las tareas de desarrollo de los adolescentes (Havighurst, 1953). De este modelo se podría deducir que en los textos de las entrevistas tendrían que aparecer relaciones jerárquicas entre categorías que enfatizan el desapego emocional del hogar paterno y categorías que representan la experiencia de la ambivalencia. Dentro de los pasajes de los textos de relaciones más estrechas con los compañeros, cabría esperar encontrar indicios de valores contradictorios por parte de los padres, pero también de los propios adolescentes, así como contradicciones entre padres y adolescentes.

9.5.2 Secuencias de determinadas categorías

Partiendo de la misma orientación hipotética, podemos suponer que en las entrevistas, tras las indicaciones de intentos de los padres de imponer sus ideas, los adolescentes informan regularmente de sus propias oposiciones. Del mismo modo, es razonable suponer que la mención de una gran libertad personal vaya seguida de un pasaje en el que se menciona la necesidad de una conexión emocional con los padres. Las secuencias inversas (= inversión de la secuencia) o las hipótesis parciales también deben comprobarse siempre. La comparación de textos en los que tales secuencias se producen según lo esperado con textos en los que se encuentra un elemento de la secuencia, pero no el otro esperado, proporciona pistas específicas del tema (p. ej., sobre las condiciones críticas del desarrollo) y teóricas (p. ej., sobre las incoherencias de la concepción del marco) especialmente importantes para análisis posteriores.

9.5.3 Agrupaciones de categorías específicas

Por ejemplo, continuando con el ejemplo anterior, se podría sospechar que las referencias a los problemas escolares, especialmente las disputas personales con profesores individuales, podrían aparecer en textos que describen el comportamiento autoritario por parte de los padres, pero no sobre su propia rebelión contra ellos. Así pues, separaríamos estas áreas (clusters) y las discutiríamos por separado. Cabe destacar una vez más que son precisamente las pruebas fallidas de este tipo de correlaciones las que pueden aportar ideas importantes para análisis posteriores. Aprendemos más de los fracasos que de los éxitos, porque entonces tenemos que reorientarnos. Los resultados negativos ni siquiera son errores, sino que muestran nuevas formas de ver el mismo material: si algo se encuentra sólo parcialmente o no se encuentra en absoluto, o incluso en un orden diferente.

Al reconstruir las conexiones sistemáticas entre las unidades de significado de los textos, se puede intentar resolver el problema de cómo encontrar estas conexiones de forma *inductiva, deductiva* o *combinando estrategias* inductivas y deductivas. Esto requiere tiempo y creatividad y no está realmente sujeto al control volitivo, ya que las percepciones tienen su propio ritmo. El enfoque preferido depende sobre todo de la pregunta de investigación. Cuando se analizan transcripciones de entrevistas no estructuradas o textos libres, como las anotaciones de un diario, se suele empezar por identificar las unidades de significado individuales y asignando los segmentos de texto correspondientes a categorías específicas. A continuación, se buscan secuencias típicas de estas categorías. Por último, se intenta asociar dichas secuencias a categorías más abstractas, el tema o los temas del hablante. Se empieza *inductivamente*, pero se pasa de las conclusiones inductivas *a las deductivas*, que funcionan de forma complementaria, de modo que se produce una cierta coherencia del sistema de análisis de forma lenta y constante.

Si uno se acerca a los datos con un interés específico por el conocimiento o una orientación teórica determinada, buscará desde el principio conexiones sistemáticas muy específicas (véase el capítulo 9.5). Se parte de hipótesis sobre posibles correlaciones y se intenta demostrarlas deductivamente a partir de los datos. Las incoherencias y los fallos dan lugar a secciones inductivas de análisis con las que se puede volver a modificar el sistema hipotético de orientación, y terminamos de nuevo con la complementariedad de la inducción y la deducción.

En el *procedimiento inductivo*, uno se esfuerza por generalizar las categorías y sus conexiones sistemáticas a partir del texto. En el *procedimiento deductivo*, se buscan pasajes concretos del texto que puedan confirmar presuposiciones generales o hipótesis sobre la conexión entre categorías. El siguiente resumen de estrategias básicas para la búsqueda sistemática de estructuras se ejecuta según este esquema. Se trata de la *búsqueda de categorías asociadas en el contexto de categorías relevantes individuales* y la *búsqueda basada en el examen de relaciones*. En función del progreso de análisis, se pueden distinguir dos subvariantes: simplicidad frente a complejidad.

9.5.4 Búsqueda en el contexto de categorías asociadas

9.5.4.1 Contexto de categorías individuales relevantes

En este enfoque, que se puede utilizar relativamente al principio del proceso de análisis, se presta atención a si la aparición de otros códigos puede registrarse antes y/o después de los segmentos de texto de una categoría crítica. Si determinados códigos aparecen con frecuencia dentro de un límite de tolerancia especificado (por ejemplo, tres líneas de texto, 20 segundos de entrevista, ..., véase también ROPE, capítulo 6.8.4.2), se puede sospechar una posible sistemática detrás de estas combinaciones. En lenguaje cotidiano, se podría formular el procedimiento de la siguiente manera: "Si aparece el código X, entonces se encuentran regularmente otros códigos en sus proximidades...". Se trata, por tanto, de una heurística con la que se puede

obtener una visión completa de las conexiones posiblemente relevantes con una determinada categoría. Por cierto, el límite de tolerancia descrito no sigue ningún criterio definido, de forma similar a ROPE en Kruschke (2017a). Esto no significa que carezca de sentido o sea completamente arbitrario. Al contrario, permite la estimación de tolerancias, y para ello es necesario un mínimo de flexibilidad. Recopilamos y anotamos estos supuestos y pruebas y creamos así una documentación lo más completa y transparente posible. Mediante diversas referencias a distintos pasajes del texto, se pueden precisar, elaborar y enriquecer estos supuestos con contenido. Todo ello sirve para responder a la pregunta de investigación.

9.5.4.2 Contexto de al menos dos categorías

La condición para la posible existencia de conexiones sistemáticas viene definida por al menos una segunda categoría a la que se debe asignar al mismo tiempo un segmento del texto. Como se describe en el capítulo 9.4, para ello se construyen tablas de codificación. La doble determinación del contenido (segmentos de texto) de las celdas de una matriz de este tipo determina la interpretación de los resultados con más fuerza que la mera determinación de una acumulación de proximidad espacio-temporal de categorías inicialmente independientes. Por otra parte, para construir una matriz de análisis es necesario un trabajo previo más conceptual, es decir, un mayor avance en el proceso de análisis. Para ello, vale la pena formular las supuestas relaciones por escrito o plasmarlas sobre el papel en forma de flechas (diagrama). Hay que tener en cuenta que dos categorías son una fuerte simplificación en comparación con la realidad. Como siempre, se trata sólo de un modelo.

9.5.5 Buscar comprobando determinadas relaciones

9.5.5.1 Comprobación de secuencias de codificación sencillas

Si, por ejemplo, de una entrevista a los padres sobre las prácticas familiares de crianza se tiene la impresión de que un padre hace un esfuerzo especial para justificar sus medidas, se podría especificar esta impresión en forma de hipótesis. En otras palabras, se buscan específicamente todos los posibles vínculos entre categorías relevantes que estén potencialmente conectadas causalmente. O supongamos, como otro ejemplo, que sospechamos conexiones sistemáticas entre acontecimientos vitales críticos y experiencias emocionales en las entrevistas biográficas. Entonces podríamos formular la siguiente hipótesis: "Cuando los entrevistados relatan acontecimientos críticos de su vida, mencionan experiencias emocionales en estrecha conexión temporal." Esto nos lleva a una búsqueda específica de los códigos correspondientes y las categorías de su conexión inmediata. Es importante obtener un sentimiento y al mismo tiempo una definición reproducible de la cercanía en el contexto del caso y aplicarla de forma coherente para el material de datos. La secuencia lógica de códigos descrita podría ser algo así

```
[Acontecimiento vital crítico y experiencias emocionales]
```

y lo asumimos dentro de, digamos, cinco líneas de texto. Esta especificación "cinco líneas de texto" se elige aquí ficticiamente y debería corresponder exactamente a la sensación para el material de datos que acabamos de mencionar, de modo que podamos seguir hablando de una proximidad temporal. A continuación buscamos sistemáticamente contraejemplos cuando se discuten acontecimientos vitales críticos pero no van seguidos de narraciones emocionales. Las narrativas emocionales siguen:

```
[Acontecimiento vital crítico y ningunas experiencias emocionales.]
```

¡La negación de la conexión entre emociones y acontecimientos vitales críticos se ha marcado con un signo de exclamación ! (= no presente). Una vez realizadas ambas búsquedas, podemos insertar los resultados como nuevo código. Este código incluye exactamente esta búsqueda y podemos etiquetarlo, por ejemplo, `SC_KritL-con-emociones` y `SC_Krit-L-sin-emociones.` Las letras `SC_` se eligen arbitrariamente y tienen por objeto garantizar la fácil localización de nuestros códigos de secuencia en el registro de códigos. No existe ninguna norma vinculante al respecto. El objetivo es simplemente obtener una visión general.

Nuestro ejemplo está orientado en nuestros hábitos. Así que ahora hemos creado codificaciones complejas y sustanciales. Si dejamos que se cuenten, detrás de cada una hay un *supuesto teórico* y *una prueba* del mismo. Un recuento, sin embargo, no implica que una mayor frecuencia vaya *siempre* acompañada de un aumento del significado. El significado se crea cuando la referencia a la pregunta de investigación se hace de la forma más clara posible en el texto. Contar puede ser una pista. Del mismo modo, una ocurrencia de k = 1 puede ser un acontecimiento fulminante, por ejemplo, cuando una interacción social da un giro inesperado.

El procedimiento descrito puede aplicarse a todas las formas de buscar y encontrar códigos complejos descritas anteriormente y a continuación.

9.5.5.2 Examinar secuencias de codificación complejas

Por ejemplo, supongamos que un investigador, leyendo entrevistas a profesores sobre sus aulas, desarrolla la hipótesis de trabajo de que, a diferencia de otros, algunos profesores describen con frecuencia el comportamiento agresivo. En relación con esto nos referimos a la falta de motivación de estos alumnos y, posteriormente, llegamos a las circunstancias familiares de los alumnos y a su consumo de juegos electrónicos. Esta hipótesis de trabajo podría y debería contrastarse con la secuencia de códigos relevantes en los textos de las entrevistas. Se espera que los resultados negativos de este ejemplo sean un indicador importante de las diferencias entre los profesores. En qué consisten estas diferencias sería objeto de un análisis posterior.

Llegados a este punto, como muy tarde, queda clara la importante contribución del software QDA al análisis de datos cualitativos. Al mismo tiempo, como ya se ha descrito, podemos asignar a las secuencias de códigos complejas su propio código en el software para hacerlo más eficiente y poder acceder a ellas directamente en el futuro. Si asignamos un nuevo código a una secuencia de códigos, este nuevo código contiene la secuencia de códigos completa, que a su vez consiste en el resultado positivo de una hipótesis sobre las conexiones de códigos en los textos. Teóricamente, los códigos creados de este modo podrían abstraerse y agregarse en *estructuras de hipótesis sobre estructuras de hipótesis*. Resulta muy problemático mantener una visión de conjunto y seguir siendo capaz de interpretar los resultados. Según nuestra experiencia, no tiene sentido en la práctica seguir vinculando dichos códigos estructurales o agregarlos en códigos aún más abstractos.

Sin embargo, no queremos excluir esta(s) posibilidad(es). Pero asignar a las secuencias de códigos su propio código sí tiene sentido. Si se dispone de unos cuantos códigos de este tipo, se pueden contarlos mediante casos y condiciones y resumirse en tablas.

A continuación, los análisis posteriores (por ejemplo, el análisis de implicados, véase el capítulo 12, o las técnicas AED, véase el capítulo 5) se basan en las estructuras, que a su vez se basan en conjeturas. De este modo, todos los resultados posteriores pueden basarse en densidades de información significativamente mayores y las afirmaciones realizadas son, en nuestra opinión, de naturaleza más sustancial en contraste con las afirmaciones basadas en categorías no relacionadas. Como resultado de las enumeraciones y otras operaciones, pueden hacer los análisis cualitativos, en principio, muy complejos y jerárquicos. No debemos olvidar que en la codificación inicial ya intervienen supuestos complejos. En el caso de los recuentos, sin embargo, no debemos ceñirnos a la cantidad como en los análisis cuantitativos, sino situar la frecuencia en un contexto significativo según el caso. Un solo acontecimiento puede ser decisivo. Si en este punto asumimos códigos más complejos, esto sólo significa *más complejos en relación con el material* del que están hechos. Se trata de una complejidad relativamente mayor, no absoluta. Teóricamente, se podría realizar esta complejidad ya durante la codificación inicial. Pero no conocemos a nadie que pueda o quiera formular hipótesis tan complejas en este momento.

9.5.6 Comparación permanente

La comparación continua de unidades de significado, categorías y variaciones de codificación dentro de un texto y entre textos diferentes constituye el núcleo de los procedimientos de análisis cualitativo y tiene su origen en la teoría fundamentada (Glaser y Strauss, 1967). Incluso en el caso de la reducción de datos dentro de un texto, no se puede lograr una codificación fiable sin comparaciones de las categorías o los segmentos de texto tanto entre sí como entre los demás textos. Esto se aplica a dos objetivos supuestamente mutuamente excluyentes: la singularidad y la generalización. Ahora bien, por un lado, en la mayoría de los estudios, se le gustaría conseguir la unicidad de cada texto más que con un sistema de categorías válido para todos los textos. La singularidad de cada texto o de la visión subjetiva del mundo expresada en él. En algún momento del proceso de investigación, por otra parte, se querría identificar configuraciones generales entre los textos. A pesar de todas las diferencias y singularidades no perder de vista lo que los textos tienen en común y cómo pueden generalizarse. Para ello hay que reconocer las conexiones sistemáticas entre los textos, compararlos y abstraer lo hallado. El objetivo aquí es acabar distinguiendo *tipos* en el material de datos.

Creamos tipologías integrando las diferencias y similitudes entre los casos y sus características. Esto se puede hacer sobre la base de análisis cuantitativos (Huber & Gürtler, 2012, cap. 12) o cualitativos (Kluge, 2000; Kelle & Kluge, 1999). Con la aplicación del álgebra de Boole a los datos cualitativos, Ragin (1987) ha desarrollado un método comparativo que se puede utilizar como procedimiento que se presta a la identificación de tipos. Según las reglas del álgebra de Boole, las constelaciones de condiciones o categorías observadas empíricamente y asignadas analíticamente a los textos se examinan con este algoritmo para ver si son pertinentes para la aparición de una categoría de referencia (= criterio). El procedimiento es reductor. El resultado son aquellas categorías (= implicantes) que tienen una conexión lógico-causal con el criterio. Todas las demás categorías sólo potencialmente relevantes se descartan por irrelevantes.

Este proceso se denomina *minimización booleana* en honor a George Boole (1814-1865), que destacó en el campo de las conexiones lógicas y el álgebra asociada ("álgebra de Boole"). Este procedimiento lógico se presenta en detalle en el capítulo 12, junto con el código R, utilizando estudios de casos empíricos.

9.6 Paradigma de codificación de estudios de caso

Para el paradigma de codificación, el análisis cuantitativo de textos (véase el capítulo 10.1) y el análisis de secuencias (véase el capítulo 11.13) utilizamos aquí el mismo material de datos o estudio de caso. Esto tiene la ventaja de poder comparar los resultados, es decir, cómo abordan un tema las distintas herramientas de investigación y qué sacan de ello. Encontrará esta comparación en el capítulo 13.5. Hemos elegido un estudio de caso bien estudiado y publicado (Studer, 1998; Gürtler, Studer & Scholz, 2012). Se trata de una carta de solicitud de una persona gravemente adicta para una plaza de terapia en régimen de internamiento. Esta carta de solicitud se envió como fax desde el centro de rehabilitación psiquiátrica a la institución de terapia de adicciones. El siguiente texto sigue el diseño original del fax:

```
[fecha]
[nombre completo del cliente]
actualmente [nombre psiquiatría].

Start again
Glärnischstr. 157
8708 Männedorf

Solicitud de entrevista
```

```
¡Hola a todos!

He leído su concepto y podía imaginar que estaría en buenas manos con
ustedes. Me tomó algún tiempo para desarrollar una perspectiva para el
futuro. Me gustaría explicárselo en una conversación personal.
Espero que me des esta oportunidad a pesar de mi edad relativamente
avanzada.

Atentamente,
[Firma nombre completo del cliente]
```

9.6.1 Cuestionario Adicción Carta de solicitud

La pregunta de investigación para los tres análisis (siguiendo el paradigma de codificación, el análisis cuantitativo de texto en el capítulo 10.1, así como el análisis secuencial en el capítulo 11.13) es la misma en cada caso:

Caso 9.1: Estudio de caso de terapia de adicciones – Formulación de la pregunta

¿Qué sabemos de la persona a partir de la carta de solicitud? ¿A qué tipo de persona, a qué características individuales nos enfrentaremos como institución cuando la invitemos a una entrevista para una terapia de adicción y qué retos podemos anticipar ya para el curso de la terapia? ¿Cuáles son los puntos fuertes y débiles del solicitante y qué podemos hacer con ellos como institución? ¿Dónde que aprender como institución y cómo podemos iniciar ya los cambios necesarios?

En concreto, no preguntamos si merece la pena ofrecer a esta persona una plaza de terapia. En primer lugar, independientemente de los resultados del análisis, sería un prejuicio totalmente inadmisible pensar de antemano sobre un cliente potencial si esta persona puede llevar una vida relativamente libre o incluso completamente libre de drogas. Los datos de que disponemos no nos permiten afirmar nada al respecto, e incluso después de una terapia exitosa o en el curso de una ruptura, esto no significa que las personas no sigan evolucionando, no importa en qué dirección.

En segundo lugar, no se trata de evaluar el potencial de éxito de la terapia, sino de reconstruir lo específico de la persona y lo que aporta a la terapia. Ni lo uno ni lo otro tienen nada que ver con el éxito en un sentido causal, aunque por supuesto ciertas características como la duración de adicción, las experiencias previas, el contexto, la situación familiar, etc. tienen una gran influencia en la persona y su potencial de recuperación. Pero una influencia siempre puede ir en tres direcciones, tanto positiva como negativa o neutra. Y la influencia puede cambiar, de positiva a negativa, viceversa o siempre hacia neutra y viceversa. Incluso con un buen análisis de una carta de solicitud, no nos atreveríamos a hacer ninguna declaración sobre el curso de la terapia y un plan de vida posterior. Los análisis de Gürtler, Studer y Scholz (2012) post-catamnésicos sobre este tema muestran de manera impresionante que los clientes encuentran soluciones a través de la vida misma a pesar de condiciones extremadamente difíciles, que *no se pueden planificar* terapéuticamente. Y lo contrario es igualmente posible, que a pesar de las mejores condiciones previas sin embargo fracasen, lo que en el caso del consumo de drogas significa una muerte prematura o, al menos parcialmente, volver a caer en los viejos patrones de comportamiento – es decir, de consumo – en las crisis.

En el capítulo 5.5.7 sobre los potenciales de recuperación en la terapia de la adicción, intentamos justificar con medios sencillos que en la muestra de Gürtler, Studer y Scholz (ibíd.), la persona con las condiciones iniciales más difíciles logró en realidad el mayor progreso relativo de autonomía. Otros antiguos clientes lograron mayores progresos absolutos según criterios sociales como la educación, el trabajo, las circunstancias familiares, etc., pero el verdadero cliente modelo, en nuestra opinión, fue el que empezó más abajo. Era aquel que, incluso años después de la terapia, seguía muy comprometido con el trabajo sobre su propia adicción. Era el que, incluso años después de la terapia, seguía muy comprometido con el trabajo sobre su propia adicción y, al mismo tiempo, saludaba a un escritor en la primera reunión con un alegre "¡Tengo un golpe en la cabeza!". Las personas son sencillamente únicas individualmente.

En la práctica organizativa, sin embargo, ya se da el caso de que cuando se llena una plaza de terapia libre, se comparan los respectivos clientes potenciales, ya que no es posible acoger a todo el mundo y además inmediatamente. Por supuesto, esta selección se basa en análisis como una carta de solicitud, una entrevista de trabajo, etc. Pero esto no significa que se dé preferencia a un cliente concreto y no significa que se dé preferencia a la persona más joven, por ejemplo, con el menor tiempo de consumo y el coeficiente intelectual más alto, por ejemplo, tenga preferencia porque supuestamente tiene más posibilidades de volver con éxito a la vida normal. Por el contrario, hay que hacer consideraciones cuidadosas sin que, afortunadamente, se pueda establecer una regla general al respecto. Éstas varían de un caso a otro.

9.6.2 Estrategia de análisis con el paradigma de codificación

Empecemos por el paradigma de codificación. Éste se caracteriza por la codificación del material, es decir, la conversión del "texto" (es decir, palabras, imágenes, audio, vídeo) en una especie de diccionario basado en el contenido, en el que los códigos representan conceptos de contenido y no necesariamente palabras individuales. Esto último, sin embargo, no es imposible, sino más bien inusual, porque las diferentes estructuras y contenidos del texto se condensan mediante la codificación, un proceso que resulta bastante difícil con palabras individuales. Las palabras individuales ya se defienden por sí mismas. A partir de los códigos creados de este modo se pueden establecer numerosas conexiones complejas entre categorías o con categorías superiores que combinan diferentes categorías y se denominan metacódigos. Las unidades estructurales de un texto, por ejemplo el cambio de orador, pero también el saludo, la introducción, el texto y la conclusión de una carta pueden estar provistas de los llamados códigos de orador, que actúan de forma simple pero eficaz como unidades estructurales y separadores. Esto permite examinar partes del texto en relación con otras partes e identificar secuencias o intercalaciones de códigos o si los códigos aparecen juntos o no, es decir, se excluyen entre sí, etc. Esto puede hacerse con una mayor cantidad de texto. Se puede complicar esto tanto como se desee con un gran número de códigos. Sin embargo, es más fácil crear metacódigos significativos y probar con ellos hipótesis más bien sencillas. No se trata de comprobar tantas hipótesis como sea posible sobre el texto, sino aquellas que resulten pertinentes para responder a la pregunta.

Es decir, podemos formular hipótesis sobre la co-ocurrencia de codificaciones de cualquier abstracción – o no – y probarlas en el texto. Si se dispone de un conjunto de datos más amplio (por ejemplo, muchas cartas de solicitud de empleo de distintas personas), se pueden realizar pruebas lógicas con la ayuda del análisis de implicaciones (véase el capítulo 12) sobre las configuraciones típicas. Por ejemplo, post-hoc a la terapia se podría examinar juntos con la carta de solicitud inicial, qué características de las cartas de solicitud en combinación con otras características del cliente o registros de la terapia hablan a favor de un curso exitoso de la terapia y cuáles no y, en caso necesario, ajustar el proceso terapéutico en consecuencia.

El reto, especialmente con el paradigma de codificación y el análisis de texto cuantitativo en este estudio de caso, es la brevedad de la carta y, por lo tanto, la pequeña cantidad de material de datos, incluso para los análisis cualitativos. La figura 9.5 muestra un extracto de la mayor parte de la codificación de la carta de solicitud, mientras que la Figura 9.6 abarca el texto asociado. Incluso el paradigma de codificación, a pesar de su orientación cualitativa, requiere en realidad una cantidad mínima de datos para poder extraer los rasgos relevantes del texto o la información. Se pueden utilizar las operaciones básicas de búsqueda y localización

de semejanzas y diferencias para elaborar las características relevantes del texto o de la información. El análisis secuencial de la Hermenéutica Objetiva, en cambio, funciona de forma secuencial, por lo que se parece más a un proceso de trabajo bayesiano: se analiza lo que está disponible. Esto es así.

Lo que no está presente es igualmente información. Se da la situación básica de razonar en un contexto de incertidumbre mínima y, en el peor de los casos, de incertidumbre máxima. Las decisiones se toman en este contexto y también en este rango. Nada de esto nos molesta porque el proceso de análisis sólo cartografía la realidad tal y como es en ese momento. Hemos experimentado en grupo que incluso es posible analizar sólo unas pocas palabras, menos que toda una frase normal en alemán. Sin embargo, esto nos permite averiguar mucho sobre la persona que pronunció la frase. La cantidad de material de datos no es despreciable en el análisis secuencial, pero sí menos relevante. Después de todo, se necesita algo para la segunda parte de la falsificación.

9.6.3 Codificación con AQUAD 7

Codificamos la presente carta según las reglas del paradigma de codificación con AQUAD 7. Suponemos que está familiarizado con la codificación y con la codificación asistida por ordenador. Nuestra estrategia para este texto es utilizar tanto el contenido como los puntos de referencia como el modo de expresión por separado. La idea es crear metacódigos y comprobar en las distintas partes de la carta la presencia de metacódigos. De ello deben derivarse hipótesis cómo se relaciona el contenido con los puntos de referencia y la forma de expresión. A partir de ahí, a su vez, obtendremos una base para derivar otras hipótesis sobre *de qué trata realmente* el escrito. Estas hipótesis sientan las bases para comprender de qué tipo de persona se trata, etc. A continuación, distinguimos entre contenido, modo, códigos del hablante y algunos puntos de referencia con el fin de dividir el texto para posteriores análisis específicos:

- El contenido corresponde a términos utilizados en la carta como **concepto, perspectiva de futuro, conversación personal, oportunidad** o **edad** y se codifican según su aparición. Dado que la carta es muy breve, hay relativamente poco contenido disponible como base de comparación, es decir, no podemos comparar cómo se utiliza el término "oportunidad" aquí y allá. Con textos más largos, este tipo de comparaciones son habituales, ya que permiten examinar el uso del mismo término en diferentes contextos. En este caso, rápidamente se hace evidente que hay muchas codificaciones individuales que, superficialmente, tienen poco que ver entre sí. Esto se debe a la brevedad del texto. La figura 9.7 da una idea de las frecuencias de los códigos. No es la única razón por la que ampliamos los contenidos simples con códigos de la forma de expresión, es decir, *cómo* se expresa un tema.

  - Conclusión Saludo
  - Dirección
  - Saludo
  - Asunto
  - Oportunidad
  - Fecha de
  - Inscripción
  - Edad
  - Referencia a "Yo"
  - Referencia a "Tú"
  - Concepto `startagain`
  - Nombre propio
  - Lugar Psiquiatría
  - Asunto
  - Entrevista en `startagain`
  - Tiempo
  - Futuro en `startagain`

Memo	desde	hasta	Código	Comienzo	longitud

Memo	von	bis	Code:	Anfang	Länge
0	1	1	Datum	0	15
0	1	3	/$Absender	0	75
0	1	3	[X] - formal	0	75
0	2	2	Name eigener	17	24
0	3	3	Ort Psychiatrie	43	32
0	7	9	/$Adresse	108	70
0	7	9	[X] - formal	108	69
0	7	9	Adresse	108	70
0	13	13	/$Betreff	211	45
0	13	13	[X] - formal	211	45
0	13	13	Subject line Betreff	211	45
0	17	17	/$Anrede	289	17
0	17	17	[S] - Betonung	289	17
0	17	17	[X] - casual	289	17
0	17	17	Anrede	289	17
0	17	17	X-unpassend	289	17
0	19	19	[S] - Imagination	359	22
0	19	19	[S] - kognitiv	326	29
0	19	19	Konzept startagain	318	37
0	19	19	X-Handlungen	359	22
0	19	19	X-kognitiv	326	29
0	19	20	/$Einstieg	318	112
0	19	20	[S] - Geborgenheit	382	47
0	19	20	[S] - Hoffnung	382	47
0	19	20	X-Emo (+)	382	47
0	19	20	X-Emo (+)	382	47
0	19	20	Zukunft in sa	356	75
0	19	24	/$Text	318	373
0	19	24	[X] - formal	318	373
0	20	20	[S] - Investition	431	29
0	20	20	X-Handlungen	431	29
0	20	21	/$Ich-Bezug	430	85
0	20	21	[S] - Planung	460	54
0	20	21	[S] - Vorzeigen von Resultaten	460	54
0	20	21	X-Handlungen	460	54
0	20	21	X-Handlungen	460	54
0	20	21	Zeit	431	84
0	21	22	/$Ihr-Bezug	515	81
0	21	22	[S] - Belehrung	514	82
0	21	22	[S] - kognitiv	515	81
0	21	22	[S] - Präsentation	515	81
0	21	22	Vorstellungsgespräch in sa	515	79
0	21	22	X-Handlungen	514	82
0	21	22	X-Handlungen	515	81
0	21	22	X-kognitiv	514	82
0	21	22	X-kognitiv	515	81
0	21	22	X-kognitiv	515	81
0	21	22	X-unpassend	514	82
0	23	23	[S] - Hoffnung	604	11
0	23	23	hohes Alter	641	34
0	23	23	X-Emo (+)	604	11
0	23	24	/$Ihr-Bezug	596	95
0	23	24	[S] - Abwehr Schutz vor Fehlschlag	615	78

Figura 9.5. *Carta de solicitud, Adicción (análisis con AQUAD 7, codificaciones)*

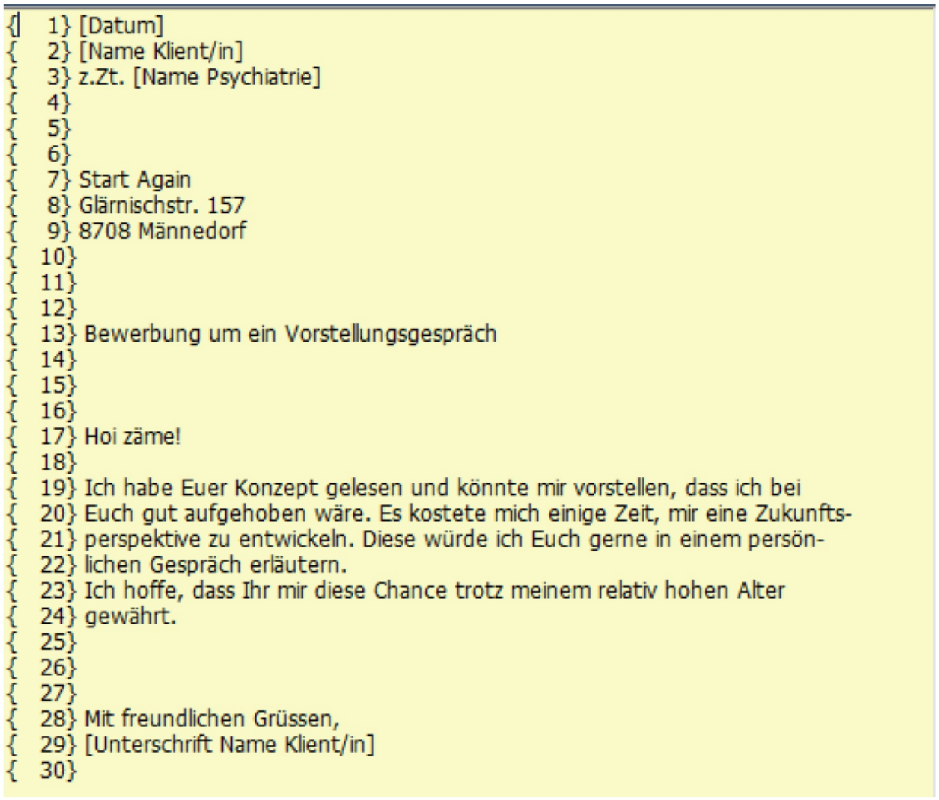


Figura 9.6. *Carta de solicitud, Adicción (análisis con AQUAD 7, texto codificado)*

- Con la forma de expresarse nos referimos, entre otras cosas, a la imaginación, la planificación, el ámbito cognitivo, la instrucción, la esperanza o la defensa (protección contra el fracaso) o la depresividad (desesperanza). Aquí ya estamos interpretando bastante para conectar los pasajes correspondientes del texto con los conceptos mencionados y otros. Estos códigos son técnicamente idénticos a los códigos de contenido, pero les anteponemos una [S] para los códigos estructurales. Esto facilita su identificación y ofrece un contraste con el contenido. Sin embargo, esto se debe principalmente a razones estructurales, para que podamos encontrar fácilmente muchas codificaciones y localizar fácilmente las codificaciones relacionadas con el contenido, estructurales, etc. Esto se debe principalmente a razones estructurales, para que siga siendo fácil orientarse en muchas codificaciones y localizar fácilmente las codificaciones relacionadas con el contenido, la estructura, etc. antes de pasar al orden alfabético. Para facilitar aún más las cosas, a continuación figuran resúmenes de los metacódigos que empiezan por X-codename.

  - [S] - Defensa Protección contra fallos
  - [S] - Criterio de exclusión
  - [S] - Instrucción
  - [S] - Énfasis
  - [S] - Depresión Desesperanza
  - [S] - Seguridad
  - [S] - Esperanza
  - [S] - Imaginación
  - [S] - Inversión
  - [S] - cognitiva
  - [S] - Planificación
  - [S] - Presentación
  - [S] - Mostrar resultados

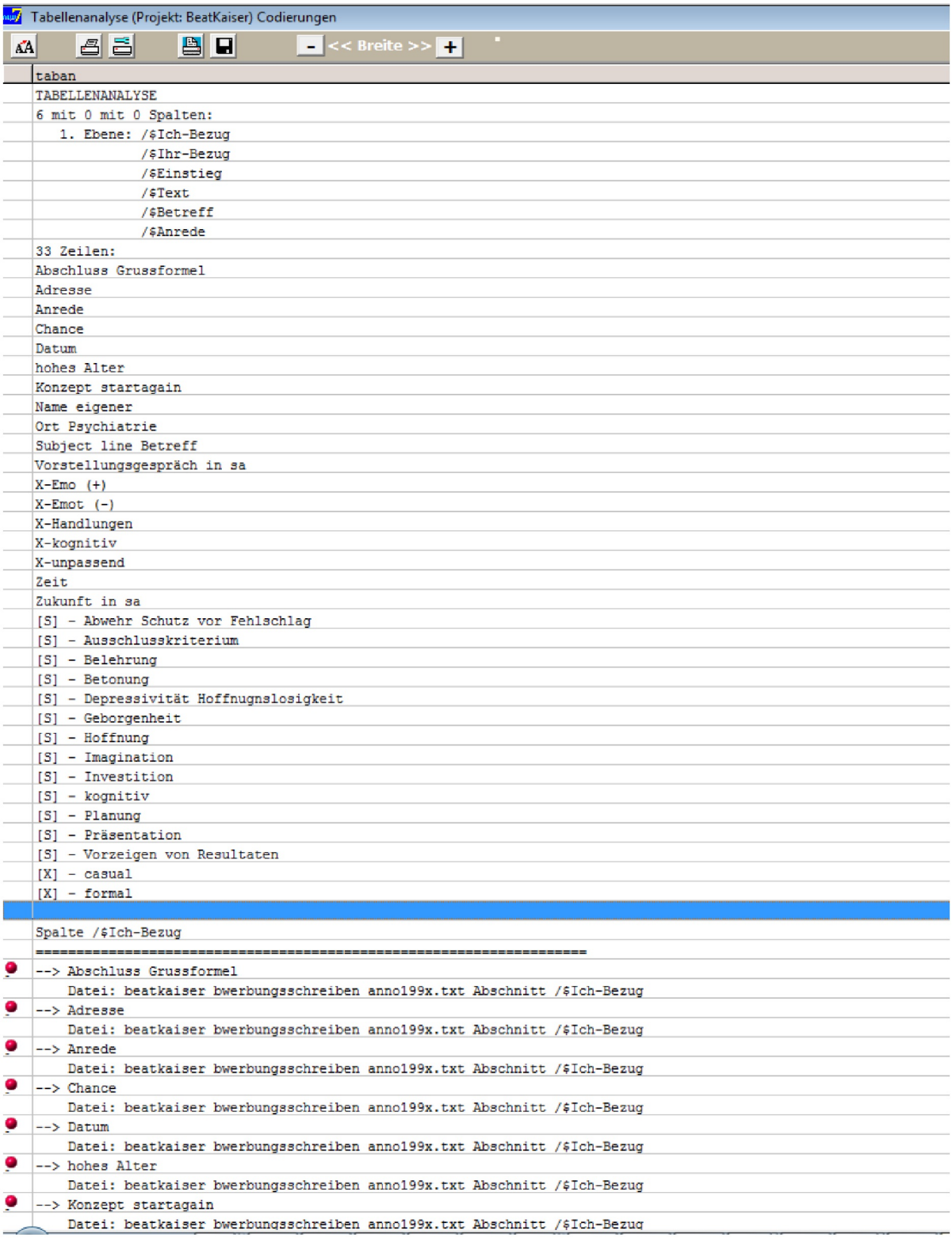


Figura 9.7. *Carta de solicitud, Adicción (análisis con AQUAD 7, frecuencias de códigos)*

- Sin embargo, esto no es suficiente, ya que tras examinar la carta se observa que hay varias unidades estructuralmente separables en el texto. Para identificar estas unidades (por ejemplo, /$dirección, /$introducción, /$texto, etc.) utilizamos los dos códigos de hablante (por ejemplo, /$dirección, /$introducción, /$texto, etc.) y hacerlas accesibles para el análisis. El resultado son los códigos de hablante

  - /$Saludo Final
  - /$Remitente
  - /$Dirección
  - /$Salutación
  - /$Asunto
  - /$Introducción
  - /$Referencia a "Yo"
  - /$Referencia a "Tú"
  - /$Texto

- Otras dos perspectivas como puntos de referencia resultan importantes: el "yo", el "mí" y el "me"- y en contraste el "tú". Utilizamos esta perspectiva específicamente para distinguir el "yo" del "tú". Utilizamos esta perspectiva para diferenciar entre los *contenidos* y el *modo* de expresarse o la posición de hablar *de aquí* frente *de allá*. Por un lado, hay un alto grado de egocentrismo en la carta, que luego siempre se contrapone a la contraparte, la institución. Se trata de un intento de comunicación, que es central debido a la brevedad de la carta. Esta figura parece significativa y debería examinarse por separado a nivel de codificación. En el transcurso del trabajo, tiene sentido denotar distintos códigos de orador para disponer de metacódigos o códigos que puedan contarse por separado más adelante y para facilitar la comprobación de hipótesis sobre el texto. Ambos códigos ya existen anteriormente como códigos de contenido. La razón es que así resulta más fácil realizar determinadas tareas de búsqueda, por ejemplo, buscar códigos de contenido dentro de un código de hablante. La figura 9.8 muestra las codificaciones con sus segmentos de archivo cada uno enumerado debajo de su código de hablante

  - /$Referencia a "Yo"
  - /$Referencia a "Tú"

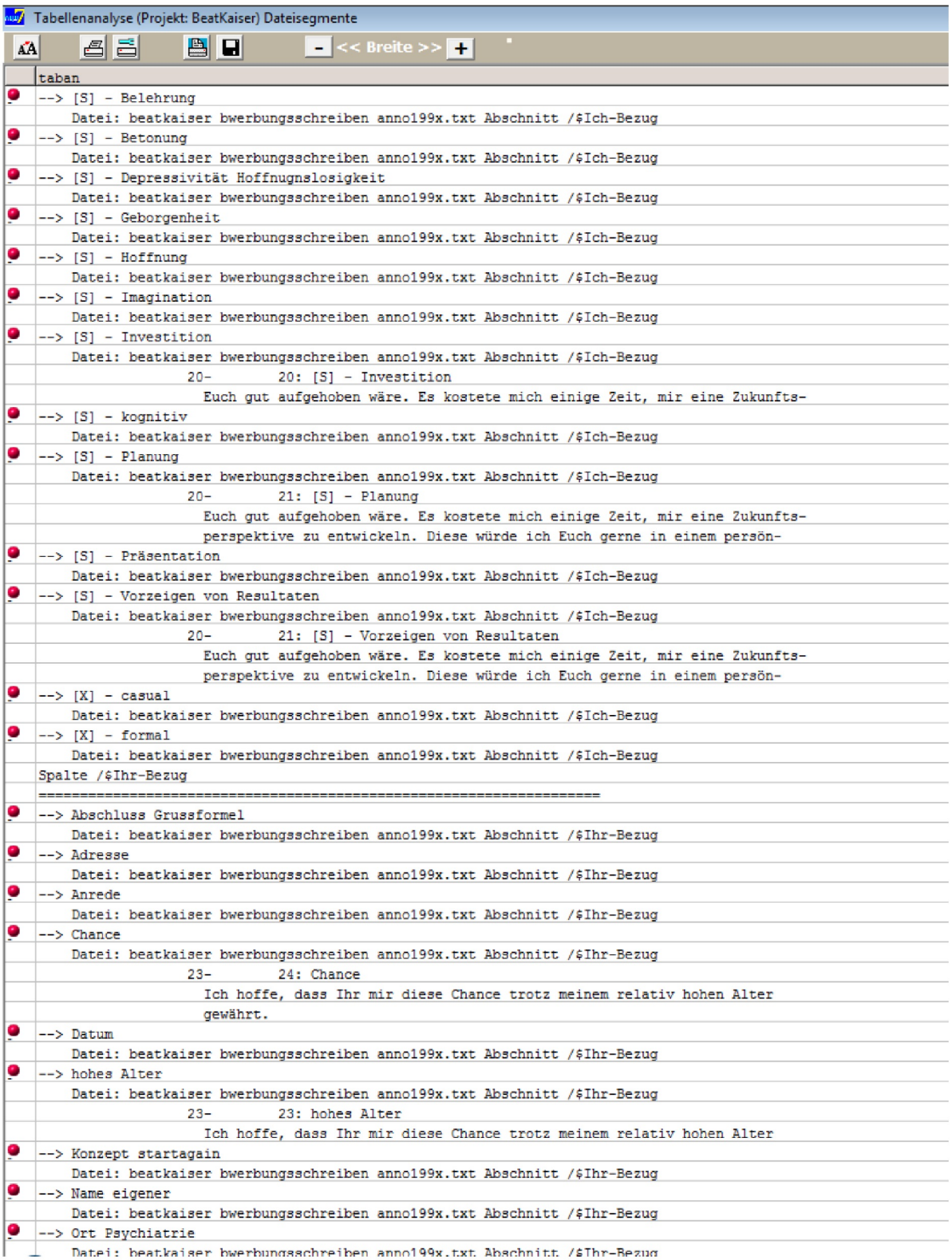


Figura 9.8. *Carta de solicitud, Adicción (análisis con AQUAD 7, codificaciones y segmentos de texto)*

9.6.4 Metacódigos

Es importante entender que los metacódigos no son mutuamente excluyentes con respecto a las codificaciones que los componen. Es decir, un código puede aparecer en diferentes metacódigos al mismo tiempo. Se trata de un comportamiento deseable, que resulta del hecho de que el significado rara vez es inequívoco, sino que puede contener diferentes interpretaciones. Sin embargo, no se debe con la disyuntividad, es decir, las codificaciones superpuestas no son mutuamente excluyentes, y esto debe tenerse en cuenta en cualquier y esto debe tenerse en cuenta en cualquier interpretación. Si se quiere evitar esto los códigos han de ser disyuntivos, de modo que excluyan todos los demás códigos de forma justificada por la regla de codificación. Como vemos, en la realidad muchos conceptos se solapan y comparten zonas comunes con otros. Si esto ocurre, lo permitimos pero somos conscientes de ello más tarde al interpretarlo y volvemos a bajar por la cadena de abstracción hasta lo concreto para rastrear las conclusiones hasta el texto.

Se pueden formar cinco metacódigos, que hemos extraído de los códigos estructurales. Como ya se ha explicado, estos códigos estructurales [S] y [X] denotan la forma en que se expresa algo en la carta de solicitud (por ejemplo, emocionalidad positiva) o qué acción está en primer plano (por ejemplo, cognitiva, planificación, etc.). El análisis demostró que estas codificaciones son más sustanciales que los temas que se abordan en términos de contenido, como el concepto de la institución, una posible entrevista de trabajo o las propias perspectivas de futuro. Estas últimas son todas menciones individuales que proporcionan una base de comparación demasiado pequeña, de modo que la necesidad de la misma codificación en contextos diferentes no emerge empíricamente. Sin embargo, los metacódigos estructurales formados son técnicamente códigos de contenido y no códigos de hablante o códigos de perfil:

- X-emotions (+) contiene códigos con contenido emocional positivo:
  - [S] - seguridad
  - [S] - esperanza
- X-emotions (-) contiene códigos con contenido emocional negativo:
  - [S] - Defensa Protección contra el fracaso
  - [S] - Depresión Desesperanza
- X-acciones contiene códigos con actividades y acciones independientes de una connotación emocional o motivacional. o connotación motivacional:
  - [S] - Instrucción
  - [S] - Imaginación
  - [S] - Inversión
  - [S] - planificación
  - [S] - presentación
  - [S] - Mostrar resultados
- X-cognitivo incluye códigos con un enfoque cognitivo en el sentido más amplio:
  - [S] - instrucción
  - [S] - cognitivo
  - [S] - presentación
- - X-inapropiado incluye codificación con contenido emocional positivo:
  - [S] - énfasis
  - [S] - instrucción

Como se puede ver, algunos de los códigos de contenido también están presentes como códigos de hablante. Como ya se ha mencionado, esto no es una contradicción siempre que el análisis se diseñe en consecuencia y los códigos no se encuentren virtualmente y, por lo tanto, se cuenten e interpreten sin obtener información. La tabla 9.6 enumera todos los códigos y metacódigos a lo largo de la subdivisión estructural

aplicada de la carta de solicitud según los códigos de hablante. Por un lado, la perspectiva en primera persona frente a otra perspectiva actúa potencialmente como códigos de hablante con /$yo-vs-tú o /$tú-vs-yo, así como el orden secuencial natural del texto según asunto, saludo, introducción y texto así como fórmula de saludo final. Remitente, dirección y cierre/saludo también están disponibles como códigos de hablante, pero no aparecen en la tabla 9.6 porque no aportan ninguna ganancia de información significativa. Utilizamos el término "ganancia de información no significativa" para describir códigos que no se salen de la norma esperada. Fuera de la norma sería el caso, si una carta de solicitud comienza sin asunto ni saludo, ya que esto contradiría la estructura de una carta formal. Sin embargo, si éste no es el caso, no prestamos más atención a la aparición de la norma esperada por el momento, siempre y cuando ésta no desempeñe un papel en el contenido de la carta – lo que da lugar a diferencias, es decir, a una ganancia de información. Entonces nos fijamos muy de cerca.

9.6.5 Codificación de secuencias

Basándonos en la tabla 9.6, hemos pensado mucho en la codificación de secuencias, pero luego nos hemos abstenido de aplicarla aquí al texto. La razón es que el texto es demasiado corto para examinar adecuadamente secuencias repartidas en distintos lugares. Eso parecería descabellado. Además, el contenido o las codificaciones contenido-estructurales entre las unidades lógicas como el saludo, la introducción, el texto o la conclusión difieren tan claramente entre sí que no tiene sentido comparar estas partes entre sí. Si dispusiéramos de más material o de un número mínimo de cartas de solicitud de distintos clientes, podríamos compararlas. En este caso examinaríamos principalmente la presencia o ausencia de codificación en el contexto de YO, TÚ y las subdivisiones de saludo, introducción y texto. Esto no tiene sentido con una sola carta.

Por ello, a efectos de demostración, la figura 9.9 muestra sólo dos pruebas de hipótesis. Aquí hemos examinado cómo se relacionan YO y TÚ [en el original: *Ich-Bezug* e *Ihr-Bezug*]. Para ello, definimos YO y TÚ como códigos de hablante y como códigos normales respectivamente y realizamos la búsqueda en AQUAD 7 con la combinación /$YO y YO frente a /$TÚ y TÚ. Sin embargo, AQUAD 7 permite la comparación de dos hablantes y ciertas combinaciones de códigos, por lo que nuestra implementación puede realizarse de diferentes maneras.

En cuanto al contenido, se trata de explorar si la referencia "tú" va seguida de una referencia "yo" o, a la inversa, si la referencia "yo" va seguida de una referencia "tú". La primera no se encontró (véase la figura 9.9 abajo), la segunda se encontró una vez (véase la figura 9.9). En el caso de una carta más larga, sería interesante investigar si se produce un nuevo cambio de perspectivas o si la visión del cliente se queda estancada en una única perspectiva (¿en algún momento?). Nuestra hipótesis explora la cuestión de cómo de flexible en la comunicación YO vs. TÚ se comporta el cliente potencial y con qué otras codificaciones se relaciona esto (por ejemplo, emocionalidad, cognición, acciones, etc.). Esto último puede examinarse observando los códigos directamente en el texto.

La figura 9.10 muestra otro aspecto en el que el código de secuencia /$YO-vs-TÚ se considera independientemente de cualquier código de hablante. En otras palabras, se pregunta en qué parte del texto aparece una RHI después de la PCI. De forma algo más amplia, se podría tomar esta estrategia para examinar si la comunicación monológica del cliente está equilibrada, es decir, si se dan tanto la perspectiva YO como la TÚ y cómo o bajo qué condiciones (es decir, códigos) se produce un posible cambio o se mantiene la perspectiva actual. La única diferencia es que esta hipótesis no se comprueba en el texto dentro de los respectivos códigos de hablante, sino independientemente de los códigos de hablante. En la presente carta, no se produce la perspectiva inversa, es decir, /$TÚ-vs-YO. Por lo tanto, no se puede mostrar un código de secuencia para ella. Otras hipótesis de secuencia podrían establecerse de forma equivalente y comprobarse en el texto y, si resultan positivas, introducirse como un código de hablante independiente y y contabilizadas posteriormente.

En este punto finalizamos la recopilación del análisis del paradigma de codificación y pasamos a los resultados.

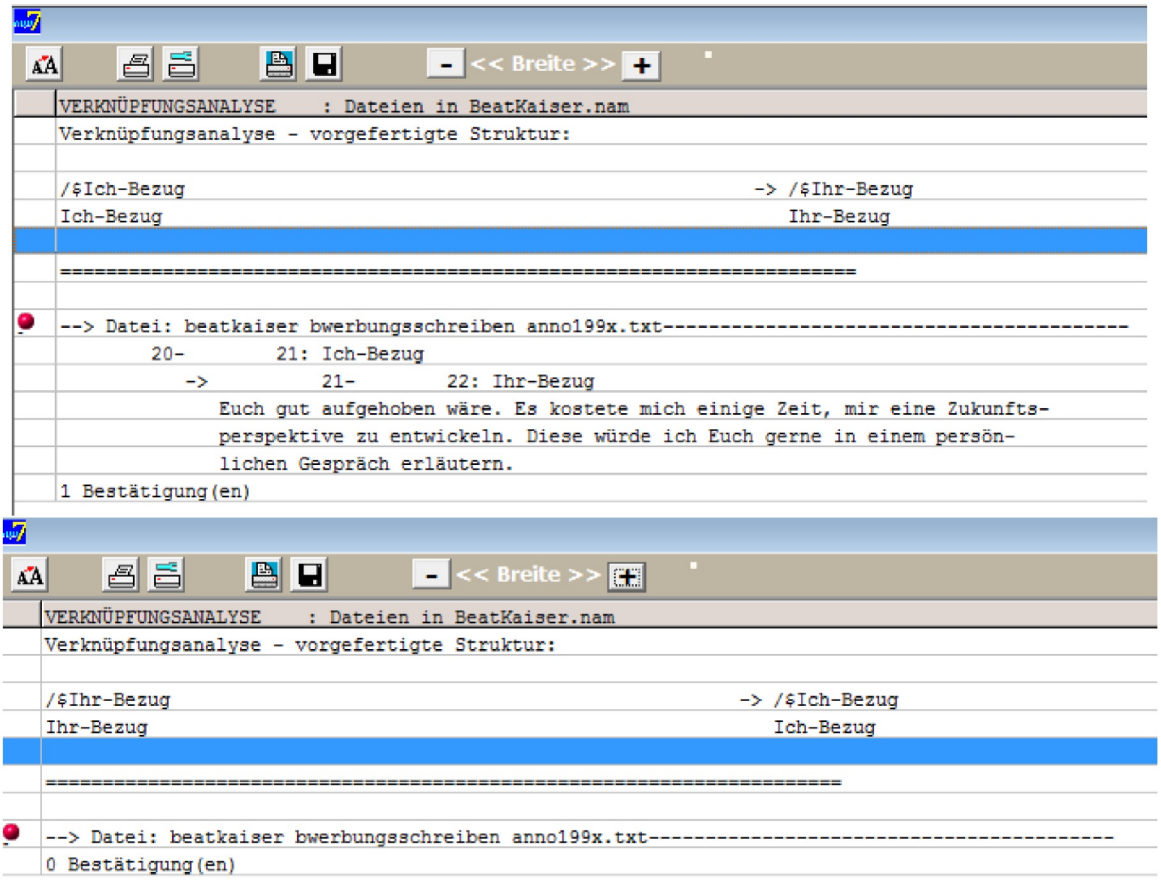


Figura 9.9. *Carta de solicitud, Adicción (análisis con AQUAD 7, codificación secuencial y segmentos de texto)*

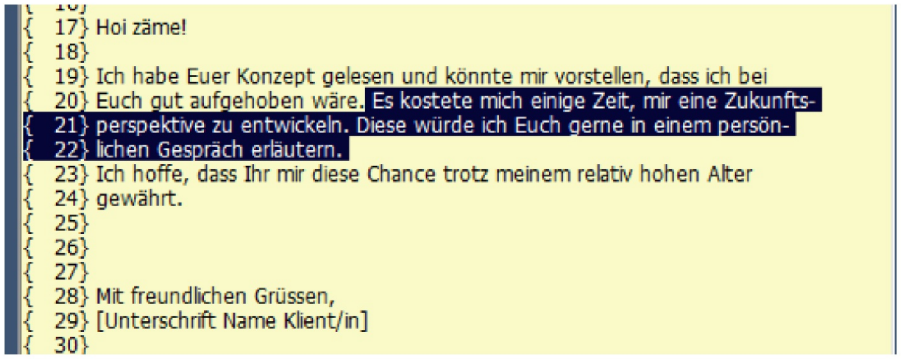


Figura 9.10. *Carta de solicitud, Adicción (análisis con AQUAD 7, yo-vs-tú codificación y segmentos de texto)*

Tabellenanalyse (Projekt: BeatKaiser) Häufigkeiten

	A	B	C	D	E	F
Abschluss Grussformel	0	0	0	0	0	0
Adresse	0	0	0	0	0	0
Anrede	0	0	0	0	0	1
Chance	0	1	0	1	0	0
Datum	0	0	0	0	0	0
hohes Alter	0	1	0	1	0	0
Konzept startagain	0	0	1	1	0	0
Name eigener	0	0	0	0	0	0
Ort Psychiatrie	0	0	0	0	0	0
Subject line Betreff	0	0	0	0	1	0
Vorstellungsgespräch in sa	0	1	0	1	0	0
X-Emo (+)	0	1	2	3	0	0
X-Emot (-)	0	2	0	2	0	0
X-Handlungen	3	2	2	6	0	0
X-kognitiv	0	3	1	4	0	0
X-unpassend	0	1	0	1	0	1
Zeit	1	0	0	1	0	0
Zukunft in sa	0	0	1	1	0	0
[S] - Abwehr Schutz vor Fehlschlag	0	1	0	1	0	0
[S] - Ausschlusskriterium	0	1	0	1	0	0
[S] - Belehrung	0	1	0	1	0	0
[S] - Betonung	0	0	0	0	0	1
[S] - Depressivität Hoffnugnslosigkeit	0	1	0	1	0	0
[S] - Geborgenheit	0	0	1	1	0	0
[S] - Hoffnung	0	1	1	2	0	0
[S] - Imagination	0	0	1	1	0	0
[S] - Investition	1	0	1	1	0	0
[S] - kognitiv	0	1	1	2	0	0
[S] - Planung	1	0	0	1	0	0
[S] - Präsentation	0	1	0	1	0	0
[S] - Vorzeigen von Resultaten	1	0	0	1	0	0
[X] - casual	0	0	0	0	0	1
[X] - formal	0	0	0	1	1	0
A: /$Ich-Bezug						
B: /$Ihr-Bezug						
C: /$Einstieg						
D: /$Text						
E: /$Betreff						
F: /$Anrede						

Figura 9.11. Carta de solicitud sobre la adicción (análisis con AQUAD 7, frecuencias de meta-/códigos por códigos de hablante)

9.6.6 Resultados descriptivos

Las figuras anteriores y las tablas 9.6 y 9.7 documentan el trabajo. Muestran extractos de la codificación (véase la Fig. 9.5), el texto durante la codificación (véase la Fig. 9.6), los códigos en el contexto del texto (véase la Fig. 9.7), los segmentos del archivo (véase la Fig. 9.8) y el recuento por frecuencia y por separado

por código de hablante (véase la Fig. 9.11). La tabla de frecuencias de los códigos separada por código de hablante (véanse las Tablas 9.6 y 9.7) proporciona la perspectiva cuantitativa dentro del enfoque cualitativo para pasar, por ejemplo, al análisis de tablas.

Pasamos así al nivel abstracto de los Metacódigos:

- X-Emo (+)
- X-Emot (-)
- X-acciones
- X-cognitivas
- X-inapropiado

En la Tabla 9.6 se observa que la emocionalidad, tanto en sentido positivo como negativo está principalmente en el contexto del TÚ y no del YO. Sin embargo, la emocionalidad atraviesa todo el texto desde el principio de la redacción. Recordemos las emociones positivas [S] - seguridad y [S] - esperanza, y las emociones negativas [S] - asco, [S] - defensa, protección contra el fracaso y [S] - depresividad. La desesperanza muestra un amplio abanico de emociones para una letra tan corta. El hecho de no poder vincularlas al (propio) YO se llamaría debilidad o vulnerabilidad emocional estructural o también inestabilidad. Esto resuena tanto con el desamparo y la falta de autonomía como con la falta de confianza en las propias capacidades. Algo así no suele surgir en la edad adulta, sino que podría ser un indicio de traumas infantiles de naturaleza inespecífica. En concreto, esto significa que el cliente podría tener problemas para experimentar su propia emocionalidad independientemente del exterior y establecer una continuidad estable, que suele ser la base de la acción racional. En la práctica, esto se pondría de manifiesto en las crisis y podría llevar al caos en la acción. La adicción sería una expresión de ello y posiblemente no la única.

El nivel de acción es más pronunciado en el YO que en el TÚ y aparece con mucha más fuerza en el curso estable de la escritura en comparación con el principio. Esto sugiere que a lo largo de una expectativa normal, el YO se presenta actuando y las ideas se desarrollan en el texto.

El nivel cognitivo se utiliza en el contexto del TÚ y, al igual que la emocionalidad, no en el contexto del YO. Asumir la responsabilidad de la propia emocionalidad y cogniciones (véase más adelante) tiene un aspecto diferente. Las frases empiezan con "encuentro/ hago/ pretendo/ etc., porque...". La cognición es mucho más prominente en el texto que al principio. Sin embargo, si contrastamos la emocionalidad con la codificación del elemento cognitivo, encontramos que [S] - cognitivo, [S] - presentación y [S] - instrucción. Es decir, la cognición no se externaliza como la emocionalidad, sino que se dirige hacia el TÚ, que se presenta y que se enseña, pero sin mostrar al mismo tiempo responsabilidad por ello. El reflejo del yo como sujeto actuante cae bastante por debajo de la mesa.

En el caso del YO, los códigos de contenido sólo muestran el tiempo como categoría. TÚ incluye la oportunidad, la vejez y la pregunta de una entrevista. Por tanto, el tiempo podría desempeñar un papel importante para el YO.

En las codificaciones estructurales de contenido [S], se observa que acciones como **planificar, invertir** y **mostrar resultados** se asocian a la YO, mientras que la TÚ abarca todas las demás codificaciones. Aquí surge la cuestión durante el análisis de que estas acciones son en gran medida de naturaleza imaginativa y no enumeran acciones o intentos reales. No se trata de una afirmación empírica sensu "he hecho", frente a la cual la contraparte tendría que posicionarse, sino que todo queda en el limbo sin crear hechos.

Tabla 9.6 *Carta de solicitud, Adicción (análisis con AQUAD 7,* frecuencias de meta-/códigos por códigos de hablante

	/$Relación a YO	/$Relación a TÚ	/$Introducción	/$Texto	/$Asunto	/$Salutación	∑
Metacódigos							
X-Emo (+)	0	1	2	3	0	0	6
X-Emot (-)	0	2	0	2	0	0	4
X-acciones	3	2	2	6	0	0	13
X-cognitivo	0	3	1	4	0	0	8
X-inadecuado	0	1	0	1	0	0	2
∑	3	9	5	16	0	0	33
Códigos de contenido							
Saludo	0	0	0	0	0	0	0
Dirección	0	0	0	0	0	0	0
Encabezamiento	0	0	0	0	0	1	1
Oportunidad	0	1	0	1	0	0	2
Fecha	0	0	0	0	0	0	0
Edad alta	0	1	0	1	0	0	2
Concepto startagain	0	0	1	1	0	0	2
Nombre propio	0	0	0	0	0	0	0
Lugar Psicatría	0	0	0	0	0	0	0
Asunto	0	0	0	0	1	0	1
Entrevista en startagain	0	1	0	1	0	0	2
Tiempo	1	0	0	1	0	0	2
Futuro en startagain	0	0	1	1	0	0	2
∑	1	3	2	6	1	1	14

Una acción real o un intento real de acción sería, por ejemplo, la mención de terapias anteriores, retiradas u otras actividades como la educación, el trabajo, la familia, etc., o incluso detalles temporales concretos sobre los propios cambios, desarrollos y planes futuros. Éstos no se mencionan. Esto podría indicar que existe una diferencia entre el pensamiento concreto y la planificación, por un lado, y la puesta en práctica, por otro. Si piensa: "¿No es el caso de todos los adictos?", se equivoca. Hay adictos que, a pesar del consumo masivo de drogas duras incluso tienen un trabajo normal, una familia, etc.. El variabilidad de los planes de vida en la adicción, como en otras áreas de la vida, es extremadamente grande y lo suficientemente grande como para que no caigamos simplemente en una imagen estereotipada generalizada.

Tabla 9.6 *Carta de solicitud, Adicción (análisis con AQUAD 7, frecuencias de meta-/códigos por códigos de hablante*

	/$Relación a YO	/$Relación a TÚ	/$Introducción	/$Texto	/$Asunto	/$Saludo	∑
Códigos estructurales del modo							
[S] - Defensa/protección contra el fracaso	0	1	0	1	0	0	2
[S] -Criterio de exclusión	0	1	0	1	0	0	2
[S] - Enseñanza	0	1	0	1	0	0	2
[S] - Ènfasis	0	0	0	0	0	0	0
[S] - Depresividad Desesperanza	0	1	0	1	0	0	2
[S] - Seguridad	0	0	1	1	0	0	2
[S] - Esperanza	0	1	1	2	0	0	4
[S] - Imaginación	0	0	1	1	0	0	2
[S] - Inversión	1	0	1	1	0	0	3
[S] - cognitivo	0	1	1	2	0	0	4
[S] - Planificación	1	0	0	1	0	0	2
[S] - Presentación	0	1	0	1	0	0	2
[S] - Mostrar resultados	1	0	0	1	0	0	2
∑	3	7	5	14	0	0	29
Códigos estructurales del modo formal de comunicación							
[X] - casual (=difuso)	0	0	0	0	0	1	1
[X] - formal (= según el rol social)	0	0	0	1	1	1	3
∑	0	0	0	1	1	2	4
∑ (total)	7	19	12	37	2	3	80

En lo que respecta a la comunicación, hay una diferencia notable entre el saludo y el texto. El saludo, por ejemplo, es difuso e inapropiado, tal y como uno se comunica realmente con los amigos o en el seno de la familia, pero desde luego no con desconocidos a nivel principalmente profesional. En cambio, el texto y el saludo se redactan posteriormente en forma de papel, es decir, en un lenguaje adecuado. Esto sugiere que el cliente es consciente de estas diferencias y utiliza intencionadamente el saludo como un "eye-catcher". como un "reclamo". Sin embargo, lingüísticamente es una infracción de las normas, que presumiblemente no se produjo sin más, sino que se utilizó deliberadamente para, de alguna manera (?), llamar la atención.

Tarea 9.1: Comprobar las hipótesis de secuencia

Basándose en la Tabla 9.6 y en las explicaciones anteriores sobre las hipótesis de secuencia del Capítulo 9.6.5, la tarea para los lectores comprometidos consistiría en examinar críticamente nuestras tesis, cuestionarlas y comprobar si también se puede llegar a lo contrario de las conclusiones anteriores. Esto exigiría o bien una ampliación de nuestras codificaciones nuevos metacódigos y posiblemente la formulación de hipótesis de secuencias o incluso una recodificación completa.

Pero ¡cuidado! - Una nueva codificación no invalida sin más los análisis existentes, porque puede ocurrir que un investigador establezca códigos diferentes, los denomine de forma distinta, etc., pero al final llegue a afirmaciones muy comparables en cuanto al contenido. Por lo tanto, una refutación de nuestro punto de vista requiere algo más que códigos con un nombre diferente, sino una buena teoría basada en el texto. Para los muy comprometidos, podría continuar de tal manera que no sólo se examine cuasi-falsificamente nuestro análisis sino que posiblemente se integre como un caso especial de una hipótesis mayor, como se sugiere en estadística: No contrastar modelos entre sí, sino establecer modelos complejos e integrar casos especiales.

Independientemente de esto, sin embargo, se puede jugar con las hipótesis secuenciales a pesar de la brevedad del texto. Por ejemplo, puede examinar cómo el YO o su referencia puede ser llevado a una construcción de hipótesis integrada con los metacódigos (emocionalidad positiva o negativa, cogniciones, acciones, ...). ¿Es realmente cierto, como hemos afirmado, que la perspectiva del YO no está relacionada con la emocionalidad (positiva, negativa) y las cogniciones, sino con el nivel de la acción? ¿Es cierto que, por el contrario, la referencia TÚ abarca todas estas cosas, pero no tanto el nivel de la acción? Dado que esto ha sido corroborado por la tabla 9.6 anterior, la pregunta más específica sería ahora formular cómo los metacódigos dependen unos de otros secuencialmente dentro o también a través de la referencia YO frente a la referencia TÚ. ¿Qué da lugar, por ejemplo de la emocionalidad positiva o negativa? ¿Es la cognición, la acción, ambas? ¿ninguna de las dos?

Tales hipótesis pueden y deben elaborarse específicamente sobre el texto.

9.6.7 Interpretación

En primer lugar, llama la atención que para una letra tan corta se hayan podido formar un gran número de códigos estructurales, es decir, que exista una gran variabilidad en el contexto reconstruible lingüísticamente. Encontramos expresiones de emocionalidad tanto positiva como negativa. Se muestran muchas actividades cognitivas, que a su vez representan aspectos motivacionales. El cliente intenta presentarse como un cliente interesante en el que merece la pena invertir. Al mismo tiempo, sin embargo, se cuelan aquí y allá elementos inapropiados para la escritura formal: por un lado, está la forma inapropiada de dirigirse a la institución terapéutica ("¡Hoi zäme!") y, por otro, la instrucción subliminal de la institución sobre cómo debe desarrollarse el programa terapéutico para el cliente según el concepto de la institución, pero basándose en las consideraciones del cliente, porque éste ha invertido tiempo. En algún momento en el futuro, aquí es donde la aspiración se encuentra con la realidad.

Si bien, por un lado, tal actitud es humanamente comprensible en cierto modo, a saber, el intento de recuperar el control sobre la propia autonomía dañada de cualquier manera, por otro lado significa que el cliente no escribe la carta siguiendo las líneas de su situación real, a saber, que necesita ayuda. Prácticamente

se sitúa un poco por encima de la acción, como un director de orquesta. Sin embargo, se encuentra exactamente en la posición opuesta, es decir, relativamente indefenso, ya que de lo contrario no necesitaría la institución y podría completar la salida de la adicción por sí mismo. Sin embargo, esto se ve contrarrestado por aspectos emocionales negativos, como si hay esperanza para él a pesar de su avanzada edad, o también el tiempo que necesitó para elaborar una perspectiva de futuro.

Tales elementos hablan de una emocionalidad débilmente apoyada e inestable que se intenta proteger. Esta diferenciación paralela de supuesta superioridad cognitiva aquí y emocionalidad débil allá caracteriza el núcleo de la carta. Podríamos derivar de ello la hipótesis de encontrarnos con una persona inteligente capaz de actuar estratégicamente, pero cuya emocionalidad es muy vulnerable o débil y está mal apoyada. Esto requerirá una gran inversión en él. Al mismo tiempo, sin embargo, no será nada fácil acercarse a esta persona, dada su edad y sus experiencias previas ciertamente traumáticas. Una hipótesis sería que la persona intenta protegerse con el intelecto y las capacidades cognitivas y puede mostrar mucha acción, planificación y posiblemente también meticulosidad o pedantería. Esto, a su vez, puede beneficiar la rutina terapéutica o entorpecerla al máximo. Probablemente depende de si el enfoque y la rutina diaria le resultan emocionalmente atractivos o no, y tal cosa es una variable bastante dinámica que – dado que la emocio-nalidad se hipotetiza débilmente – puede cambiar casi cada hora. Así que lo que la persona necesita es el desarrollo de una emocionalidad interna estable autorreferencial con un uso simultáneo dirigido de lo cognitivo, para que ambos aspectos (emoción, cognición) tiren juntos y apunten en la misma dirección y no tengan un efecto desintegrador. Consideramos que la motivación en sí es pronunciada y muy elevada. Tanto la reflexión sobre la propia edad como la expresión de inseguridad al respecto en el plano emocional sugieren que nos encontramos con un cliente motivado cuyas acciones, sin embargo, a veces pueden ir en una dirección y luego en otra o incluso salir mal. En el contexto de un centro de terapia de adicciones basado en el principio de una comunidad residencial, esto significa problemas a nivel de la interacción social, es decir, el trato con otros clientes y el personal y el manejo general de reglas, órdenes, acuerdos, metas, objetivos, etc. Estos son probablemente los retos con este cliente. La referencia de la emocionalidad a uno mismo y, por tanto, independizarse de los demás sería otro importante para estabilizar la emocionalidad debilitada y la fragilidad. Actualmente la emocionalidad debilitada controla indirectamente el intelecto presumiblemente bien desarrollado. Esto provoca discrepancias en la práctica, ya que una crisis emocional no está limitada por la cognición, sino que presumiblemente se intensifica, ya que ésta aparece subordinada a lo emocional. Lo emocional sería, por tanto, el lugar donde debe tener lugar una postsocialización separada.

9.6.8 Conclusión

Si resumimos todas las partes del análisis, esperaríamos encontrarnos con una persona que tiene capacidades cognitivas y estratégicas atractivamente altas y, al mismo tiempo, una emocionalidad débilmente apoyada e inherentemente inestable. Existe una alta motivación para la terapia debido a la elevada edad y a las correspondientes experiencias previas. En principio, la persona podría tener sus propios puntos de vista sobre la aplicación e interpretación de los procedimientos y contenidos en la terapia cotidiana, el concepto y las normas institucionales necesarias, lo que podría suponer un reto adicional a la hora de tratar con ella.

Dejaremos en este punto el análisis. Este debe considerarse como preliminar y sirve de demostración y no pretende ser completo. Cualquiera que desee analizarlo más a fondo es bienvenido a hacerlo. Se adjuntan los archivos del proyecto AQUAD 7, incluida la carta de solicitud, junto con los scripts de R.

Capítulo 10

El Análisis de Contenidos Manifiestos – el Análisis Cuantitativo de Textos

> "Ten menos curiosidad por las personas y más curiosidad por las ideas".
>
> Marie Curie, 1867-1934

En el capítulo 9.1 nos remitimos a la distinción entre contenido latente (Kracauer, 1952) y manifiesto (Berelson, 1952) para diferenciar el paradigma de codificación del análisis de datos no-numéricos. En el capítulo 9.2, analizamos variantes de la búsqueda de significados ocultos y latentes en los datos, es decir, en el caso del material textual, formas de *leer entre líneas*. En cambio, el análisis cuantitativo de textos consiste en el análisis sistemático del contenido manifiesto, es decir, el análisis de lo que, según Berelson, está escrito en *blanco y negro* en las líneas y en la superficie. por la forma que también aquí, en algún momento del proceso de análisis, se presuponen y se extraen conclusiones que apuntan a contenidos latentes. Sólo el camino y la información utilizada difieren de las explicaciones anteriores.

La técnica central del análisis cuantitativo de textos es la determinación de las características del contenido manifiesto. En el caso más sencillo, se cuentan las palabras individuales de un texto y se enumeran según su calidad. Contando todas las palabras y comparando sus aptitudes en los textos, se pueden obtener al menos unas primeras indicaciones de los principales puntos de significado. Para esto último, sin embargo, son necesarias suposiciones cualitativo-interpretativas adicionales: *¿en qué palabras se pueden considerar una indicación de qué?*, o *¿qué indica la ausencia de determinadas palabras en un texto?* En principio, esto no dice nada sobre la estructura y el significado de estas palabras contadas. Podríamos contar la letra "e" en la lengua alemana y, tras compararla con otras letras, nos daríamos cuenta de que se trata de la letra más común del alfabeto alemán. Sin embargo, de ello no podemos extraer ninguna conclusión sobre el contenido.

En muchas aplicaciones del análisis cuantitativo de textos, la frecuencia de cada palabra contenida en el texto no se determina simplemente de forma global, sino que el recuento va precedido de una selección cualitativa-interpretativa de palabras críticas. *Crítico* se define siempre como un indicador de contenido en el sentido de la pregunta de investigación. Esto requiere un intenso trabajo previo.

No podemos entrar en la polémica sobre el uso de enfoques cuantitativos en el análisis de datos cualitativos. No hay duda, sin embargo, de que para muchas preguntas de investigación es útiles complementar los análisis cualitativos con los cuantitativos (véase el capítulo 13). Si se dispone de palabras clave como indicadores de determinadas afirmaciones en los textos, entonces se pueden obtener al menos unas primeras indicaciones sobre los focos de significación contando estas palabras y comparando posteriormente sus frecuencias en relación con los textos estudiados. Otros analistas textuales irían más lejos y utilizarían los resultados de un análisis de frecuencias a nivel de palabras como datos de partida para procedimientos estadísticos pertinentes, como los análisis de clusteres. A continuación, se pueden realizar análisis de combinaciones de palabras. Vorderer y Groeben (1987) ofrecen un análisis detallado de las frecuencias de palabras. Este ámbito, el PNL ("procesamiento del lenguaje natural") y los análisis

lexicométricos de textos (Tristl, Müller & Bachmann, 2015) por lingüistas siguen casi sin problemas, pero mucho más orientados estadísticamente y sobre todo menos preocupados por el contenido y sus realidades relacionales.

Dado que las palabras individuales pueden tener a menudo significados diferentes, dependiendo del contexto en el que aparezcan en los textos analizados, no se pueden descartar errores en el recuento de palabras clave, ya que no se tienen en cuenta los diferentes contextos de significado. Un procedimiento útil para garantizar que sólo se tengan en cuenta aquellas apariciones de una palabra clave para las que realmente la palabra crítica tiene la función indicadora especificada es la elaboración de listas KWIC ("KeyWord In Context") o listas de concordancia (Weber, 1990), que muestran la palabra crítica en el contexto del texto. Supongamos que en las redacciones de los alumnos sobre el uso de los juegos de ordenador, queremos utilizar la palabra clave "juego" como indicador para saber qué piensan personalmente sobre estos juegos. Una lista KWIC – reproducida aquí sólo parcialmente para un único ensayo – tendría entonces este aspecto:

- "*Juego* a veces por la noche con mi ..."
- "No *juego* mucho, ya que salgo durante el día."
- "Creo que también *juego* con éxito"
- "Mis padres dicen que está bien que no *juego* mucho ...".

El texto original dice:

"...y sólo se mejora si se sigue practicando. Creo que también juego con éxito, porque siempre hay quien es mejor y quien empieza de nuevo. Mis padres dicen que está bien que no *juego* estos juegos y no me paso horas en ello ..."

Aquí queda inmediatamente claro que el "juego" de la última frase de la lista KWIC no señala la opinión del escritor, sino la de sus padres. Y en la segunda frase, "juego" no es suficiente, porque es combinado con "también". El control del significado mediante la inserción de las palabras clave en su contexto lingüístico también está disponible si se utiliza un programa informático para buscar simplemente las palabras clave seleccionadas en el conjunto del texto y marcarlas en color. El análisis cuantitativo de textos y la PNL pueden hacer uso hoy en día de todos los algoritmos y toda la potencia de los ordenadores. Sin embargo, se necesita un buen modelo para la transición de las frecuencias y los análisis al significado, que es una tarea difícil, sobre todo en casos individuales. La búsqueda de significado, por otra parte, es el único objetivo del siguiente capítulo sobre el paradigma de la reconstrucción, que se centra exclusivamente en la reconstrucción objetiva del significado y contextos de significado.

10.1 Estudio de caso: Análisis cuantitativo de textos

Contrastamos el estudio de caso del capítulo 9.6, que se analizará con más detalle en el capítulo 11.13, con un análisis de texto cuantitativo exploratorio. Se hace hincapié en lo de exploratorio porque no tenemos ninguna hipótesis inferencial basada en el contenido para probar en el texto. El reto ahora es que apenas tenemos material, porque la carta de solicitud es corta, un fax. Además, contiene elementos formales como remitente, dirección, asunto, saludo y salutación final, todo lo cual proporciona sólo una pequeña parte de la información que podría ser importante para nosotros. El propio texto de la solicitud es breve.

Respecto a `R` - Silge y Robinson () han publicado un libro gratuito sobre minería de textos con R, que se basa en la infraestructura extendida del paquete R `tidytext` y paquetes R asociados como `tm, dplyr, broom, tidyr` y `ggplot2`. Más información se encuentra en Welbers, van Atteveldt y Benoit (2017). Además, como de costumbre, están las excelentes viñetas, que suelen formar parte de los paquetes de R. En R, hay paquetes para el análisis cuantitativo de textos como `tm` (minería de textos), `SnowballC` (stemming

de textos), `wordcloud` (generador de nubes de palabras), `quanteda` (gestión y análisis de textos), `koRpus` (análisis de textos), `lsa` (análisis semántico latente) y `OpenNLP` (colección de herramientas de procesamiento del lenguaje natural). Para el análisis de redes sociales como Twitter, Reddit, Youtube, etc., existen paquetes especiales de R como `vosonSML, tuber` y `tidytext` o `ggwordcloud` (Spörlein, 2021). Las operaciones con cadenas son posibles con los paquetes R `stringi, stringr, readr, textclean, textshape` y otras funciones R ya disponibles en las funciones básicas de R (por ejemplo, `substr(), grep(), gsub(), gregexpr(), strsplit(), paste(), nchar(), tolower(), toupper(), ...`). Básicamente, éstas son las funciones básicas para modificar textos – buscar y reemplazar, localizar y contar, eliminar, dividir y combinar y fusionar, y conversiones como mayúsculas y minúsculas. Importante es la funcionalidad de usar expresiones regulares y alguna sintaxis de Perl. Para ello, existen viñetas propias como en el paquete de R `stringr`. Esto proporciona potentes herramientas para preparar textos para el análisis. Los paquetes de R `sentimentr` y `lexicon` también ofrecen léxicos y el análisis cuantitativo del discurso `qdap`, que tiene su propio sistema de gestión de textos (Rinker, 2021).

En cuanto a la importación, los paquetes de R pertinentes ofrecen la importación no sólo desde archivos de texto puro. Es posible incluir fuentes de Internet, así como hojas de cálculo Excel (R Core Team, 2021) con los paquetes de R `xlsx` o `readxl u openxlsx`, así como mediante `readtext` formatos basados en texto como `html, json, xml, pdf` así como `odt, doc, docx` o `rtf` – por nombrar sólo algunos. Los paquetes de R enumerados más arriba también ofrecen la posibilidad de importación de redes sociales.

10.1.1 Importación de textos

Empezamos leyendo el texto (`ptIII_qual_quan-textanalysis.r`)

```
# leer texto código R
stextnam <- c("Bewerbungsbrief_BK_sa.txt")
stext <- readLines(stextnam)
stext
```

y comprobamos la importación con

```
> # check whether import was ok
> stext
[1] [fecha]
[2] [nombre_cliente]
[3] actualmente [nombre_psiquiatría]
[4]
[5]
[6] Empezar de nuevo
[7] Glärnischstr. 157
[8] 8708 Männedorf
[9]
[10]
[11]
[12] Solicitud de entrevista
[13]
[14]
[15] ¡Hoi zäme!
[16]
[17]
[18] He leído vuestro concepto y podía imaginar que estaría en buenas manos con vosotros.
[19] Me llevó algún tiempo pensar en una perspectiva
[20] de futuro. Me gustaría explicártelo en una
[21] conversación personal.
[22] Espero que me dé esta oportunidad a pesar de mi edad
[23] relativamente avanzada.
```

```
[24]
[25]
[26] Atentamente,
[27]
[28] [Firma Nombre_Cliente]
```

Incluimos la versión alemana, que era la base de los análisis siguentes:

```
[1] "[Datum]"
[2] "[Name KlientIn]"
[3] "z.Zt. [Name_Psychiatrie]"
[4] ""
[5] ""
[6] "Start Again"
[7] "Glärnischstr. 157"
[8] "8708 Männedorf"
[9] ""
[10] ""
[11] ""
[12] "Bewerbung um ein Vorstellungsgespräch"
[13] ""
[14] " "
[15] "Hoi zäme!"
[16] ""
[17] ""
[18] "Ich habe Euer Konzept gelesen und könnte mir vorstellen, dass ich bei"
[19] "Euch gut aufgehoben wäre. Es kostete mich einige Zeit, mir eine Zukunfts-"
[20] "perspektive zu entwickeln. Diese würde ich Euch gerne in einem persön-"
[21] "lichen Gespräch erläutern."
[22] "Ich hoffe, dass Ihr mir diese Chance trotz meinem relativ hohen Alter"
[23] "gewährt."
[24] ""
[25] ""
[26] "Mit freundlichen Grüssen,"
[27] ""
[28] "[Unterschrift Name_KlientIn]"
```

Obviamente el texto tiene 28 filas:

```
> # number of rows
> length(stext)
[1] 28
```

10.1.2 Preparación del texto y trabajo preliminar para los análisis

Dado que el fax original es el mismo que la conversión informática en cuanto a longitudes de línea, saltos, etc., echamos un vistazo a las longitudes de línea. Antes de eso, descomponemos el texto en sus palabras, signos de puntuación, espacios, etc., es decir, la unidad más pequeña posible.

```
> # split text into single chars including everything
> stext.cut <- sapply(stext, function(x) strsplit(x,"",fixed=TRUE))
> head(stext.cut)
$`[Datum]`
[1] "[" "D" "a" "t" "u" "m" "]"
$`[Name KlientIn]`
[1] "[" "N" "a" "m" "e" " " "K" "l" "i" "e" "n" "t" "I" "n" "]"
$`z.Zt. [Name_Psychiatrie]`
[1] "z" "." "Z" "t" "." " " "[" "N" "a" "m" "e" "_" "P" "s" "y"
[16] "c" "h" "i" "a" "t" "r" "i" "e" "]"
[[4]]
```

```
character(0)
[[5]]
character(0)
$‘Start Again‘
[1] "S" "t" "a" "r" "t" " " "A" "g" "a" "i" "n"
```

Ahora contamos las palabras

```
> # count words
> stext.wfreq <- stri_count_words(stext)
> stext.wfreq
[1] 1 2 2 0 0 2 2 2 0 0 0 4 0 0 2 0 0 12 12 11
[21] 3 12 1 0 0 3 0 2
```

y los caracteres individuales por línea

```
> # count characters per line including empty spaces
> s.text.lbyrow <- unlist(lapply(stext.cut, length))
> names(s.text.lbyrow) <- 1:length(s.text.lbyrow)
> s.text.lbyrow
 1  2  3  4  5  6  7  8 9 10 11 12 13 14 15 16 17 18 19 20 21
 7 15 24  0  0 11 17 14 0  0  0 37  0  1  9  0  0 69 73 70 26
22 23 24 25 26 27 28
69  8  0  0 25  0 28
```

e insertamos ambos en un diagrama de barras (véase la Fig. 10.1).

```
names(stext.wfreq) <- names(s.text.lbyrow)
# plot
par(mfrow=c(2,1))
barplot(stext.wfreq, col="green", pre.plot=grid(),
main="Words per line",
xlab="Line", ylab="Frequency")
barplot(s.text.lbyrow, col="skyblue",
pre.plot=grid(), main="Characters per line",
xlab="Line", ylab="Frequency")
```

Otro paso sería dividir el texto en palabras sueltas y, paralelamente, limpiarlo. El objetivo es hacer directamente accesible la unidad más pequeña del texto. Existen diferentes estrategias: En primer lugar, podemos reducir el texto a formas troncales. Utilizamos el algoritmo de Porter (1980), que se utiliza específicamente para distintas lenguas. El se hace con `wordStem()` de `SnowballC` o `wordStem()` de `tm`:

```
# stem languages
getStemLanguages()
wordStem(stext, language="german")
stemDocument(stext, language="german")
```

Si comparamos las soluciones, observamos que no son idénticas:

```
> # comparison
> wordStem(stext, language="german") == stemDocument(stext,
+ language="german")
[1]   TRUE FALSE TRUE  TRUE TRUE  TRUE TRUE  TRUE  TRUE  TRUE
[11]  TRUE FALSE TRUE FALSE TRUE  TRUE TRUE FALSE FALSE FALSE
[21] FALSE FALSE TRUE  TRUE TRUE FALSE TRUE  TRUE
```

Sería necesario examinar cómo funciona `wordStem()` en comparación con `stemDocument()`.

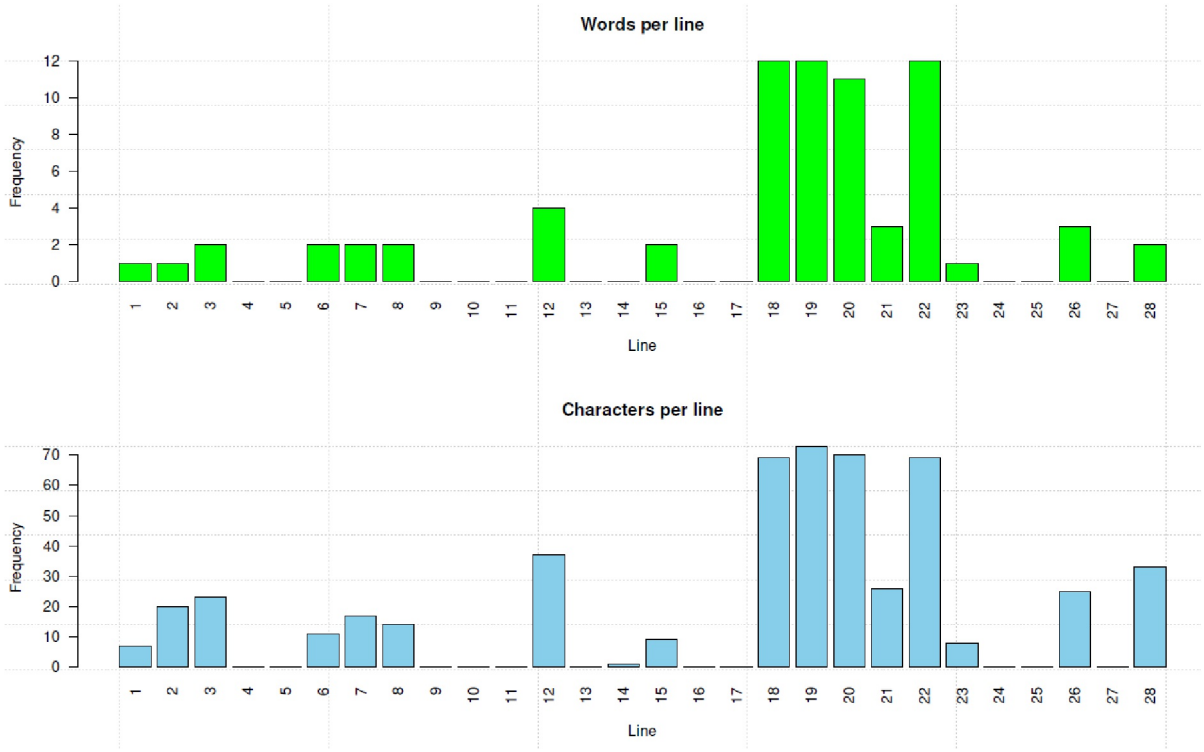


Figura 10.1. *Carta de solicitud – adicción*
(gráficos de barras, palabras o caracteres individuales por línea)

Un punto de partida general para el análisis de textos lo ofrecen `tm` o `quanteda`. Cuando se trabaja con muchos documentos, `tm` o `quanteda` y la mayoría de los demás entornos de procesamiento del lenguaje natural (PLN) empiezan leyendo el texto como un corpus. Un corpus es una colección de textos. La función de R `corpus()` en `tm` o en `quanteda` permite leer diferentes textos y proporciona una infraestructura con la que pueden trabajar las demás funciones del paquete. Del término corpus hay que distinguirlo de la matriz de términos del documento (DTM). Se trata de una matriz numérica que contiene las frecuencias de los términos que aparecen en los textos. Las columnas representan los documentos y las filas los términos. Las celdas contienen las frecuencias.

Lo mostramos con el ejemplo de `tm`:

```
>#  read as a corpus
> corps <- Corpus(VectorSource(stext))
> corps
< <SimpleCorpus> >
Metadata: corpus specific: 1, document level (indexed): 0
Content: documents: 28
```

Podemos ver si el documento se ha leído correctamente con

```
# check document
inspect(corps)
```

Podemos insertar separadores en los lugares de todos los signos de puntuación y dividir así el texto sólo en palabras. Sin embargo, esto tiene el inconveniente de que no podemos registrar la longitud media de las frases, las partes de las frases, etc. Tampoco queda claro cuándo se utilizan puntos, comas, signos de exclamación, signos de interrogación, etc. También es posible que estos signos tengan un significado para nosotros con el fin de responder a nuestra pregunta sobre el contenido. Todo esto es información potencialmente relevante que no podemos perder ni desechar tan fácilmente. En un análisis visual del fax

original puede ser importante fijarse en el espaciado entre caracteres, palabras, secciones, etc. etc. Con un texto escrito en un ordenador, no solemos tener esos problemas. Además, sólo tenemos un conocimiento limitado de cómo se creó el texto. No toda la información es accesible para el análisis, parte se nos escapa. Asimismo, cuando pudimos separar los caracteres especiales y/o molestos al leer en el texto. Nuestro texto no tenía ningún carácter especial, pero añadimos algunos al principio para demostrarlo:

```
> # remove special characters
> stext.alt <- readLines("Bewerbungsbrief_BK_sa_w-spec-char.txt")
> head(stext.alt,3)
[1] "@@\\\\[Datum]"
[2] "[Name KlientIn]"
[3] "z.Zt. [Name_Psychiatrie]"
```

Ahora pasamos al procedimiento basado en expresiones regulares para eliminar, entre otros, los caracteres especiales. Los sustituimos por espacios:

```
# remove special characters
stext.alt <- readLines("Bewerbungsbrief_BK_sa_w-spec-char.txt")
head(stext.alt,3)
# convert to corpus
corps.alt <- Corpus(VectorSource(stext.alt))
inspect(corps.alt)
# we use regular expressions but not perl style
rem.pattern <- content_transformer(
function(x, pattern) gsub(pattern, " ",
x, perl=FALSE, fixed=FALSE))
inspect( tm_map(corps.alt, rem.pattern, "/") )
inspect( tm_map(corps.alt, rem.pattern, "@") )
inspect( tm_map(corps.alt, rem.pattern, "[\\]") )
# =
inspect( tm_map(corps.alt, rem.pattern, "\\\\") )
rem.pattern2 <- content_transformer(
function(x, pattern) gsub(pattern, " ",
x, perl=FALSE, fixed=TRUE))
# =
inspect( tm_map(corps.alt, rem.pattern2, "\\") )
```

Para facilitar la comparación, podría cambiar todo a minúsculas o eliminar muchas cosas como números, espacios múltiples, signos de puntuación, ciertas palabras basadas en diccionarios comunes basadas en diccionarios específicos de lenguas comunes o creados especialmente, etc. Todos estos deseos podemos realizar con funciones de R como `tolower()`, `removeNumbers()`, `stripWhitespace()`, `removePunctuation()`, `removeWords()`, etc. Echamos un vistazo a esto, pero sólo de forma limitada, ya que no tiene mayor relevancia para nosotros en este caso. Con textos más largos y textos de orígenes diferentes, sin embargo, esto puede tener un aspecto muy diferente. Debería quedar claro que introducir cambios en el texto tiene consecuencias y modifica la base de información.

```
# remove "unnecessary" parts from the text
# lower cases
tolower(stext)
# remove numbers
removeNumbers(stext)
# remove white space
stripWhitespace(stext)
# remove punctuations
removePunctuation(stext)
# remove stop words for a specific language
removeWords(stext, stopwords("German"))
# remove one's own stop words
mystopws <- c("ich", "Ich", "mir", "mich")
removeWords(stext, mystopws)
# =
```

```
mystopws2 <- c("ich", "mir", "mich")
removeWords(tolower(stext), mystopws2)
```

Empezaremos dividiendo el texto donde haya espacios. Antes de empezar, nos referimos a la posibilidad de operaciones anidadas en R. Desde el ámbito Linux/Unix, esto se conoce como el símbolo de Pipe "|". Permite utilizar la salida de un proceso como entrada de otro. Algo de Bash muestra esto – lo llamamos desde R con `system()`. Lo que se pide en la carta de solicitud son todos los lugares donde se menciona "yo" (o "mi", "me", etc. La salida de la lista positiva se ordena alfabéticamente por la primera letra de la línea:

```
> system("cat Bewerbungsbrief_BK_sa.txt | grep 'ich' | sort")
Euch gut aufgehoben wäre. Es kostete mich einige Zeit, mir eine Zukunfts-
Ich habe Euer Konzept gelesen und könnte mir vorstellen, dass ich bei
lichen Gespräch erläutern.
Mit freundlichen Grüssen,
perspektive zu entwickeln. Diese würde ich Euch gerne in einem persön-
```

Eso todavía no es muy exacto. Vemos la aparición de "lichen" en la línea 3 o "freundlichen" en la línea 4. No estamos buscando eso. Pero el significado del símbolo de la Pipe debería haber quedado claro. R ya permite tal procedimiento en parte debido a su orientación a objetos:

```
# object orientation instead of pipe
hist(replicate(100, mean(rnorm(100))))
```

El paquete R `magrittr` amplía los anidamientos e introduce el operador %>%. El paquete R `quanteda` para la minería de textos se basa en gran medida en este enfoque. Algo comparable puede encontrarse en el campo gráfico con el paquete R `ggplot2`. Apliquémoslo.

Dado que los signos de puntuación no desempeñan ningún papel en lo que sigue, podemos eliminarlos. Primero transformamos todo a minúsculas, eliminamos los números y los espacios múltiples, así como los signos de puntuación. Luego dividimos el texto en los espacios y eliminamos las entradas vacías resultantes.

```
# split text and remove everything that
# is not a word, term, verb, whatever
stext.red <- tolower(stext) %>% removeNumbers() %>%
stripWhitespace() %>% removePunctuation()
# split text at empty space
stext.red.wonly <- unlist(strsplit(stext.red, " "))
# remove empty entries
empty.IDs <- stext.red.wonly == ""
stext.red.wonly <- stext.red.wonly[!empty.IDs]
```

Ahora miramos el resultado:

```
> stext.red.wonly
[1] "datum"                  "name"                   "klientin"
[4] "zzt"                    "namepsychiatrie"        "start"
[7] "again"                  "glärnischstr"           "männedorf"
[10] "bewerbung"             "um"                     "ein"
[13] "vorstellungsgespräch"  "hoi"                    "zäme"
[16] "ich"                   "habe"                   "euer"
[19] "konzept"               "gelesen"                "und"
[22] "könnte"                "mir"                    "vorstellen"
[25] "dass"                  "ich"                    "bei"
[28] "euch"                  "gut"                    "aufgehoben"
[31] "wäre"                  "es"                     "kostete"
[34] "mich"                  "einige"                 "zeit"
[37] "mir"                   "eine"                   "zukunfts"
[40] "perspektive"           "zu"                     "entwickeln"
[43] "diese"                 "würde"                  "ich"
[46] "euch"                  "gerne"                  "in"
[49] "einem"                 "persön"                 "lichen"
```

```
[52] "gespräch"          "erläutern"          "ich"
[55] "hoffe"             "dass"               "ihr"
[58] "mir"               "diese"              "chance"
[61] "trotz"             "meinem"             "relativ"
[64] "hohen"             "alter"              "gewährt"
[67] "mit"               "freundlichen"       "grüssen"
[70] "unterschrift"      "nameklientin"
```

Básicamente, ya queda bastante bien. Lo que se nota es que los saltos de línea originales no se han tratado bien. Por ejemplo, la palabra separada "persönlich", representada en el texto como "persön-" y "lich", se ha dividido y ahora da lugar a dos entradas. Tendríamos que limpiar esto de antemano con un esfuerzo manual o con un algoritmo inteligente antes de dividir nada. Para ello, habría que buscar los saltos de línea con guión y comprobar, a partir del léxico, si se trata simplemente de un salto de línea con división de palabras. En este caso, habría que volver a unir las palabras, siempre que el salto de línea no tenga ningún significado en términos de contenido. Este no debería ser el caso en la mayoría de los textos. Otra posibilidad es tratar de identificar todos los signos de puntuación para separarlos, pero conservándolos al mismo tiempo. Éste sería el caso, por ejemplo, si no se tratara sólo de una cuestión de palabras, secuencias, proximidad y distancia, etc. Se podrían ver los signos de puntuación como parte del texto y contarlos o analizarlos después. No hemos podido encontrar una función de R para ello. El procedimiento es sencillo. Empezamos por definir un conjunto de caracteres que podemos utilizar como separadores.

```
# split text but maintain
# use a special character to split
# ! " # $ % & ' ( ) * + , - . / : ; < = > ? @ [ \ ] ^ _ ` { | } ~
# we use only some punctuations from the document itself
puncts <- c(".",",","!","?","-")
```

A continuación, insertamos espacios antes y después de estos signos de puntuación

```
stext.split <- stext
for(i in puncts)
{
stext.split <- gsub(i, paste(" ",i," ",sep=""),
stext.split, fixed=TRUE)
}
```

dividimos en los espacios en blanco y eliminamos las entradas vacías con

```
# actual split R-Code
stext.split <- unlist(strsplit(stext.split," ", fixed=TRUE))
# remove empty entries
stext.split <- stext.split[!stext.split == ""]
stext.split
```

Lo que resulta es:

```
 [1] "[Datum]"            "[vollerNameKlientIn]"          "z" R-Output
 [4] "."                  "Zt"                            "."
 [7] "[NamePsychiatrie]"  "Start"                         "Again"
[10] "Glärnischstr"       "."                             "157"
[13] "8708"               "Männedorf"                     "Bewerbung"
[16] "um"                 "ein"                           "Vorstellungsgespräch"
[19] "Hoi"                "zäme"                          "!"
[22] "Ich"                "habe"                          "Euer"
[25] "Konzept"            "gelesen"                       "und"
[28] "könnte"             "mir"                           "vorstellen"
[31] ","                  "dass"                          "ich"
[34] "bei"                "Euch"                          "gut"
[37] "aufgehoben"         "wäre"                          "."
[40] "Es"                 "kostete"                       "mich"
```

```
[43] "einige"           "Zeit"                 ","
[46] "mir"              "eine"                 "Zukunfts"
[49] "-"                "perspektive"          "zu"
[52] "entwickeln"       "."                    "Diese"
[55] "würde"            "ich"                  "Euch"
[58] "gerne"            "in"                   "einem"
[61] "persön"           "-"                    "lichen"
[64] "Gespräch"         "erläutern"            "."
[67] "Ich"              "hoffe"                ","
[70] "dass"             "Ihr"                  "mir"
[73] "diese"            "Chance"               "trotz"
[76] "meinem"           "relativ"              "hohen"
[79] "Alter"            "gewährt"              "."
[82] "Mit"              "freundlichen"         "Grüssen"
[85] ","                "[Unterschrift"        "vollerNameKlientIn]"
```

Eso ya tiene mejor pinta. En realidad ahora tendríamos que eliminar cosas superfluas como `vollerNameKlientIn]` o `[Unterschrift`. Es importante tener en cuenta que faltan la corchete a la izquierda de `vollerNameKlientIn]` y la correspondiente corchete de cierre a la derecha de `[Unterschrift`. Podemos utilizar estos términos, ya que no son necesarios para el cliente potencial, podemos dejarlos en el texto. Son sólo marcadores de posición para preservar el anonimato con respecto al uso de nombres reales. Eliminamos los corchetes.

```
# remove "[" and "]" R-Code
puntcs2 <- c("[","]")
puntcs2
for(i in puntcs2) stext.split <- gsub(i, "", stext.split, fixed=TRUE)
stext.split
```

Como se puede ver, gran parte del análisis de textos consiste en prepararlos para satisfacer las propias aspiraciones de investigación. Hay que tomar una serie de decisiones, todas ellas en torno a qué material entra en el análisis propiamente dicho y qué se elimina de antemano.

10.1.3 Frecuencias de palabras

Ahora podemos empezar a contar. Antes de hacerlo, vamos a darnos una visión gráfica del texto. Para ello son populares las nubes de palabras. Éstas se pueden crear con `wordcloud` del paquete `wordcloud` de R. La tarea consiste primero en contar las palabras y luego presentarlas gráficamente (véase la Fig. 10.2 a la izquierda)

```
# word cloud
set.seed(1234)
# count words
stext.split.wfreq <- table(tolower(stext.split))
stext.split.wfreq
# plot
wordcloud(words=names(stext.split.wfreq),
        freq=stext.split.wfreq,
        min.freq=1, max.words=200,
        random.order=FALSE, rot.per=0.35,
        colors=brewer.pal(8, "Paired"))
```

y, a continuación, las palabras y caracteres individuales (véase la fig. 10.2):

```
# single characters
# count characters and signs
stext.cut.wfreqtab <- table(unlist(stext.cut))
stext.cut.wfreqtab
```

```
# plot
wordcloud(words=names(stext.cut.wfreqtab),
        freq=stext.cut.wfreqtab,
        min.freq=1, random.order=FALSE, rot.per=0.35,
        colors=brewer.pal(8, "Dark2"))
```

Hay varias estrategias para contar. Una sería limpiar el texto, contarlo y trazar sólo las frecuencias que superen un determinado criterio (véase la Fig. 10.3).

```
# plot word freqs
stext.tkized <- tokens(stext.split,
        what="word",
        remove_numbers=TRUE,
        remove_punct=TRUE,
        remove_symbols=TRUE,
        remove_separators=TRUE,
        remove_url=TRUE) %>%
#tolower() %>%
tokens_remove(pattern=words("vollernameklientin",
        "unterschrift",
        "namepsychiatrie"))
stext.tkized.tab <- sort(table(tolower(unlist(stext.tkized))),
        dec=TRUE)
critfreq <- 1
stext.tkized.tab.cut <- stext.tkized.tab[stext.tkized.tab > critfreq]
stext.tkized.tab.cut
barplot(stext.tkized.tab.cut, las=2,
        names.arg=names(stext.tkized.tab.cut),
        col="green", main="Word frequencies",
        ylab="freq")
```

Figura 10.2. *Carta de solicitud – adicción (nubes de palabras, frecuencias de palabras y caracteres individuales)*

Volvemos al objeto `corps`. Ahora no nos limitamos a contar por palabras, sino que utilizamos un procedimiento del paquete `tm` de R creando una matriz de documentos

```
> # build freq matrix
> dtm <- TermDocumentMatrix(corps)
> dtm.mat <- as.matrix(dtm)
> head(dtm.mat)
                  Docs
Terms             1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
[datum]           1 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
[name             0 1 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
klientin]         0 1 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
[name_psychiatrie] 0 0 1 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
z.zt.             0 0 1 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
again             0 0 0 0 0 1 0 0 0  0  0  0  0  0  0  0  0  0
Docs
Terms             19 20 21 22 23 24 25 26 27 28
[datum]            0  0  0  0  0  0  0  0  0  0
[name              0  0  0  0  0  0  0  0  0  0
klientin]          0  0  0  0  0  0  0  0  0  0
[name_psychiatrie]  0  0  0  0  0  0  0  0  0  0
z.zt.              0  0  0  0  0  0  0  0  0  0
again              0  0  0  0  0  0  0  0  0  0
```

y contar sobre ellos

```
dtm.mat.freq <- data.frame(freq=sort(rowSums(dtm.mat), decreasing=TRUE))
```

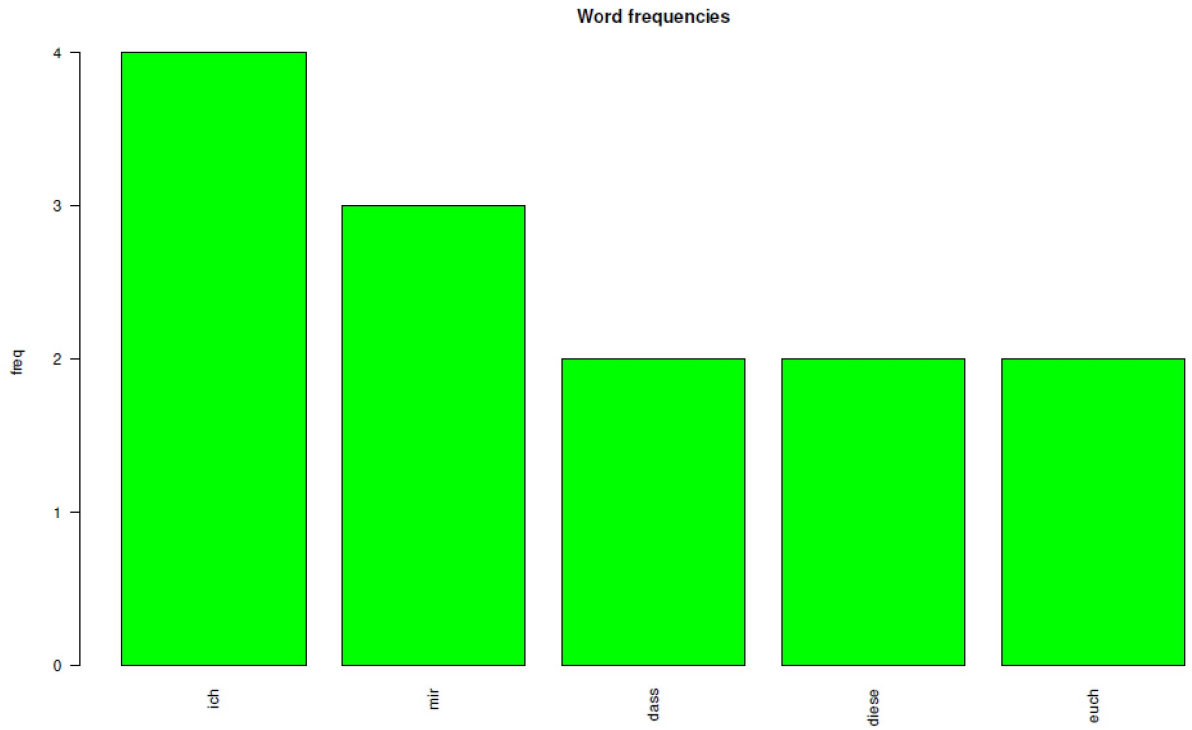


Figura 10.3. Carta de solicitud - Adicción (Gráfico de barras, frecuencias de palabras)

La tabla de frecuencias `dtm.mat.freq` tiene entonces este aspecto – hay principalmente entradas con frecuencias singulares. Esta última requiere su propia interpretación:

	freq
ich	4
mir	3
dass	2
euch	2
diese	2
[datum]	1
[name	1
klientin]	1
[name_psychiatrie]	1
z.zt.	1
again	1
start	1
157	1
glärnischstr.	1
8708	1
männedorf	1
bewerbung	1
ein	1
vorstellungsgespräch	1
hoi	1
zäme!	1
bei	1
euer	1
gelesen	1
habe	1
konzept	1
könnte	1
und	1
vorstellen,	1
aufgehoben	1
eine	1
einige	1
gut	1
kostete	1
mich	1
wäre.	1
zeit,	1
zukunfts-	1
einem	1
entwickeln.	1
gerne	1
perspektive	1
persön-	1
würde	1
erläutern.	1
gespräch	1
lichen	1
alter	1
chance	1
hoffe,	1
hohen	1
ihr	1
meinem	1
relativ	1
trotz	1
gewährt.	1
freundlichen	1
grüssen,	1
mit	1
[unterschrift	1
name_klientin]	1

La tabla de frecuencias refleja lo que ya sabemos por las nubes de palabras. Las palabras más frecuentes se refieren al "yo" y a los demás con "yo", "me" y "tú", así como a las estructuras que indican procesos con

"eso" y a los nombres concretos con "estos". Simplificado en pocas palabras: alguien tiene en mente algo concreto con los demás.

10.1.4 Sensibilidad al contexto – palabras clave contextualizadas

A partir de la tabla de frecuencias, se plantea la cuestión de en qué contexto (antes, después) se sitúa una palabra. Una posibilidad consiste en crear un diccionario de términos relevantes y buscarlos. Esta estrategia KWIC (= palabra clave en contexto) es un método sencillo y popular para obtener una impresión exploratoria de los contextos. También se practica en el análisis de datos cualitativos asistido por ordenador (véase el capítulo 11.12).

Si ampliamos esta idea de proximidad frente a distancia de los contextos relevantes, esto puede tener sus trampas en la práctica. Así, de forma comparable al paradigma de codificación, podemos (¡o deberíamos!) definir dentro de qué rango específico del texto los términos siguen mostrando una cercanía y en qué punto muestran una forma de distancia. Conceptualmente, esto se corresponde con el concepto ROPE de la estadística bayesiana o la cuestión de dónde deben trazarse los umbrales de significación en la estadística clásica, si es que se quiere trabajar con dichos umbrales críticos. La cuestión del umbral crítico de significación y de información en el sentido de Bateson (1985) es una cuestión cualitativa original.

Aquí nos interesa la perspectiva del cliente potencial y examinamos la perspectiva del yo y, posteriormente, la opuesta al tú/usted/vosotros. Por ejemplo, podemos definir un pequeño diccionario y, en el sentido de KWIC (= palabra clave en contexto), investigar en qué contexto de más/menos tres o cuatro palabras se sitúan nuestras palabras clave. Para ello utilizamos el conjunto de datos `stext.split` creado anteriormente. Las palabras clave son todas las palabras con una frecuencia mayor que uno. Definimos el rango como más/menos cuatro pasos, es decir palabras, antes y después de la palabra clave.

```
# KWIC
stext.split
# dictionary with words freq > 1
KWICdict <- rownames(dtm.mat.freq)[dtm.mat.freq > 1]
KWICdict
# where do we find a keyword?
stext.split.L <- tolower(stext.split)
kwic.IDs <- lapply(seq_along(KWICdict),
       function(x) which(stext.split.L == KWICdict[x]))
names(kwic.IDs) <- KWICdict
kwic.IDs
stext.split.l <- length(stext.split)
steps <- 4
#
kwic.res <- sapply(kwic.IDs, function(x)
{
 xl <- length(x)
 sapply(seq_along(x), function(i)
 {
  number <- x[i]
  start <- ifelse(number < 1,1,number-steps)
  end <- ifelse(number > stext.split.l,stext.split.l,number+steps)
  stext.split[start:end]
 })
})
```

Veamos el resultado `kwic.res`. Cada entrada debe leerse de arriba abajo. En el centro está siempre la palabra clave.

```
$ich
     [,1]                    [,2]          [,3]                [,4]
[1,] "Vorstellungsgespräch"  "mir"         "entwickeln"        "lichen"
[2,] "Hoi"                   "vorstellen"  "."                 "Gespräch"
```

```
[3,] "zäme"                 ","             "Diese"              "erläutern"
[4,] "!"                    "dass"          "würde"              "."
[5,] "Ich"                  "ich"           "ich"                "Ich"
[6,] "habe"                 "bei"           "Euch"               "hoffe"
[7,] "Euer"                 "Euch"          "gerne"              ","
[8,] "Konzept"              "gut"           "in"                 "dass"
[9,] "gelesen"              "aufgehoben"    "einem"              "Ihr"
$mir
     [,1]                   [,2]            [,3]
[1,] "Konzept"              "mich"          "hoffe"
[2,] "gelesen"              "einige"        ","
[3,] "und"                  "Zeit"          "dass"
[4,] "könnte"               ","             "Ihr"
[5,] "mir"                  "mir"           "mir"
[6,] "vorstellen"           "eine"          "diese"
[7,] ","                    "Zukunfts"      "Chance"
[8,] "dass"                 "-"             "trotz"
[9,] "ich"                  "perspektive"   "meinem"
$dass
     [,1]                   [,2]
[1,] "könnte"               "."
[2,] "mir"                  "Ich"
[3,] "vorstellen"           "hoffe"
[4,] ","                    ","
[5,] "dass"                 "dass"
[6,] "ich"                  "Ihr"
[7,] "bei"                  "mir"
[8,] "Euch"                 "diese"
[9,] "gut"                  "Chance"
$euch
     [,1]                   [,2]
[1,] ","                    "."
[2,] "dass"                 "Diese"
[3,] "ich"                  "würde"
[4,] "bei"                  "ich"
[5,] "Euch"                 "Euch"
[6,] "gut"                  "gerne"
[7,] "aufgehoben"           "in"
[8,] "wäre"                 "einem"
[9,] "."                    "persön"
$diese
     [,1]                   [,2]
[1,] "perspektive"          ","
[2,] "zu"                   "dass"
[3,] "entwickeln"           "Ihr"
[4,] "."                    "mir"
[5,] "Diese"                "diese"
[6,] "würde"                "Chance"
[7,] "ich"                  "trotz"
[8,] "Euch"                 "meinem"
[9,] "gerne"                "relativ"
```

Otra estrategia para realizar KWIC sería con el paquete R quanteda:

```
> # KWIC again
> kwic.dict <- c("ich","dass")
> stext.merged <- do.call("paste", as.list(stext.split, sep=" "))
> stext.merged
[1] "Datum vollerNameKlientIn z . Zt . NamePsychiatrie Start Again
Glärnischstr . 157 8708 Männedorf Bewerbung um ein Vorstellungsgespräch
Hoi zäme ! Ich habe Euer Konzept gelesen und könnte mir vorstellen ,
dass ich bei Euch gut aufgehoben wäre . Es kostete mich einige Zeit ,
mir eine Zukunfts - perspektive zu entwickeln . Diese würde ich Euch
gerne in einem persön - lichen Gespräch erläutern . Ich hoffe , dass Ihr
mir diese Chance trotz meinem relativ hohen Alter gewährt . Mit
```

```
freundlichen Grüssen , Unterschrift vollerNameKlientIn"
> corps3 <- tokens(stext.merged)
> kwic(corps3, pattern=phrase(kwic.dict))
Keyword-in-context with 6 matches.
[text1, 22]    ein Vorstellungsgespräch Hoi zäme! | Ich |
           habe Euer Konzept gelesen und
[text1, 32]    und könnte mir vorstellen, | dass |
           ich bei Euch gut aufgehoben
[text1, 33]    könnte mir vorstellen, dass | ich |
           bei Euch gut aufgehoben wäre
[text1, 56]    zu entwickeln. Diese würde | ich |
           Euch gerne in einem persön
[text1, 67]    - lichen Gespräch erläutern. | Ich |
           hoffe, dass Ihr mir
[text1, 70]    erläutern. Ich hoffe, | dass |
           Ihr mir diese Chance trotz
```

Aquí, el texto completamente dividido se empaquetó en una sola cadena, se convirtió en un objeto tokens y se analizó con kwic(). En primer lugar, definimos nuestro propio diccionario KWIC, en este caso con las palabras Ich y dass (yo y que).

La búsqueda de asociaciones puede realizarse con findAssocs() del paquete tm de R. Elegimos r = 0.4 como límite inferior de correlación. Nos limitamos a las palabras con frecuencias superiores a 1, preparamos el conjunto de datos y eliminamos sólo ciertas palabras, pero no las stop-words (palabras de parada). Las stop-words son palabras específicas de un idioma que son más bien triviales y tienen poco significado, como "the" en inglés, que corresponde a "ein[e]" en alemán. Los paquetes de R stopwords y tm contienen cada uno la función de R stopwords(), que se puede utilizar de forma específica para cada idioma. Lógicamente, la utilizaríamos con respecto al idioma alemán, aunque el saludo de la carta de solicitud esté en alemán de Suiza. Se echa en falta una biblioteca de palabras de parada para los dialectos. Además, el resto de la carta está en alto alemán, por lo que el uso local puede resultar significativo. Sin embargo, como todas las palabras nos parecen importantes por el momento, no eliminamos ninguna. Obtenemos las palabras de parada específicas de cada país con stopwords() y podemos eliminarlas con tokens_remove():

```
# remove stop words R-Code
swGerman <- stopwords("german")
swGerman
length(swGerman) #= 231
tokens_remove(tokens(stext.split), swGerman)
```

Se puede modificar el siguiente código R ligeramente para eliminar estas palabras de todos modos. Asimismo, las palabras pueden reducirse a sus tallos, que también se desactivan a continuación. Aplicamos el procedimiento

```
# set a lower limit for frequencies R-Code
# frequent terms and associations
corps2 <- Corpus(VectorSource(stext))
inspect(corps2)
# remove stuff
corps2 <- tm_map(corps2, tolower) %>%
tm_map(removeWords, words("vollernameklientin",
       "unterschrift",
       "namepsychiatrie")) %>%
tm_map(removePunctuation) %>%
tm_map(removeNumbers)
inspect(corps2)
dtm2 <- TermDocumentMatrix(corps2)
dtm2.mat <- as.matrix(dtm2)
VIPterms <- findFreqTerms(dtm2, lowfreq=2)
VIPterms
# find associations for VIP terms
for(i in VIPterms)
{
```

```
        print( findAssocs(dtm2, terms=i, corlimit=0.43) )
}
```

y miramos el resultado:

```
$dass
       ich    mir  bei      euer  gelesen       habe
      0.85   0.80 0.69      0.69     0.69       0.69
   konzept könnte  und vorstellen    alter    change
      0.69   0.69 0.69       0.69     0.69       0.69
     hoffe  hohen  ihr     meinem  relativ     trotz
      0.69   0.69 0.69       0.69     0.69       0.69
     diese
      0.46
$ich
      dass    bei       euer gelesen  habe konzept
      0.85   0.81       0.81    0.81  0.81    0.81
    könnte    und vorstellen     mir diese
      0.81   0.81       0.81    0.67  0.54
$mir
      dass     ich     bei       euer    gelesen  habe
      0.80    0.67    0.56       0.56       0.56  0.56
   konzept  könnte     und vorstellen aufgehoben  eine
      0.56    0.56    0.56       0.56       0.56  0,56
    einige    gut  kostete       mich       wäre  zeit
      0.56   0.56     0.56       0.56       0.56  0.56
  zukunfts   alter  chance      hoffe      hohen   ihr
      0.56    0.56    0.56       0.56       0.56  0.56
    meinem relativ   trotz
      0.56    0.56    0.56
$euch
   aufgehoben   eine   einige   gut    kostete  mich
         0.69   0.69     0.69  0.69       0.69  0.69
         wäre   zeit zukunfts einem entwickeln gerne
         0.69   0.69     0.69  0.69       0.69  0.69
  perspektive persön    würde diese
        0.69    0.69     0.69  0.46
$diese
    einem entwickeln gerne perspektive persön  würde
     0.69       0.69  0.69        0.69   0.69   0.69
    alter     chance hoffe       hohen    ihr meinem
     0.69       0.69  0.69        0.69   0.69   0.69
  relativ      trotz   ich        dass   euch
     0.69       0.69  0.54        0.46   0.46
```

Si tenemos muchos textos, vale la pena tener una visión tabular de las palabras de interés. Definamos de nuevo un pequeño diccionario y apliquémoslo en nuestro caso a un solo texto.

```
> # count terms over many documents
> dict.obj <- dictionary(list(ich = c("ich","mir"),
        + euch=c("euch","euer"),
        + begr=c("diese","dass")
        + ))
> dfm(tokens_lookup(corps3, dict.obj,
        valuetype="glob", verbose=TRUE))
applying a dictionary consisting of 3 keys
Document-feature matrix of: 1 document,
        3 features (0.00% sparse) and 0 docvars.
features
docs    ich euch begr
  text1 7   3    4
```

10.1.5 Colocaciones

Otra visión interesante de un texto es la de las colocaciones, es decir, la co-ocurrencia de palabras en en proximidad directa. Entonces sólo tenemos que justificar cómo "cerca" es exactamente denota. Puede tratarse de una vecindad directa o de una más lejana, como dentro de más/menos x palabras. Aquí, a efectos de demostración, realizamos toda una serie de transformaciones: Eliminación de números, signos de puntuación, símbolos y separadores, así como una lista de palabras vacías del alemán y la aplicación del algoritmo de Porter para la raíz de las palabras (Porter, 1980) antes de pasar a los N-gramas. Todo ello en una única función R autoescrita `ngram()`.

```
# collocations R-Code
# define stop words language specific
swGerman <- stopwords("german")
ngram <-function(txt, nofngram=2, ntop=10,
        REMsw=TRUE, wlistrem=swGerman)
{
    tokens(txt,
        what="word",
        remove_numbers=TRUE,
        remove_punct=TRUE,
        remove_symbols=TRUE,
        remove_separators=TRUE) %>%
    tokens_remove(pattern=wlistrem) %>%
    tokens_wordstem() %>%
    tokens_ngrams(n=nofngram) %>%
    dfm() %>%
    topfeatures(n=ntop)
}
```

y probarlos para diferentes longitudes de n-gramas con y sin distancia de palabras de parada típicas:

```
> ngram(stext,3,10)
konzept_gelesen_vorstellen          gelesen_vorstellen_dass
1                                   1
gut_aufgehoben_wäre                 aufgehoben_wäre_kostet
1                                   1
wäre_kostet_zeit                    kostet_zeit_zukunfts-
1                                   1
hoff_dass_chanc                     perspekt_entwickeln_gern
1                                   1
entwickeln_gern_persön-             dass_chanc_trotz
1                                   1
> ngram(stext,3,10, wlistrem="")
bewerbung_um_ein                    um_ein_vorstellungsgespräch
1                                   1
ich_habe_euer                       habe_euer_konzept
1                                   1
euer_konzept_gelesen                konzept_gelesen_und
1                                   1
gelesen_und_könnte                  und_könnte_mir
1                                   1
könnte_mir_vorstellen               mir_vorstellen_dass
1                                   1
> ngram(stext,4,10)
konzept_gelesen_vorstellen_dass     gut_aufgehoben_wäre_kostet
1                                   1
aufgehoben_wäre_kostet_zeit         wäre_kostet_zeit_zukunfts-
1                                   1
perspekt_entwickeln_gern_persön-    hoff_dass_chanc_trotz
1                                   1
dass_chanc_trotz_relativ            chanc_trotz_relativ_hohen
```

```
1                                    1
trotz_relativ_hohen_alter
1
> ngram(stext,4,10, wlistrem="")
bewerbung_um_ein_vorstellungsgespräch ich_habe_euer_konzept
1                                    1
habe_euer_konzept_gelesen            euer_konzept_gelesen_und
1                                    1
konzept_gelesen_und_könnte           perspekt_zu_entwickeln_dies
1                                    1
gelesen_und_könnte_mir               und_könnte_mir_vorstellen
1                                    1
könnte_mir_vorstellen_dass           zu_entwickeln_dies_würde
1                                    1
> ngram(stext,5,10)
gut_aufgehoben_wäre_kostet_zeit
1
aufgehoben_wäre_kostet_zeit_zukunfts-
1
hoff_dass_chanc_trotz_relativ
1
dass_chanc_trotz_relativ_hohen
1
chanc_trotz_relativ_hohen_alter
1
> ngram(stext,5,10, wlistrem="")
ich_habe_euer_konzept_gelesen        habe_euer_konzept_gelesen_und
1                                    1
euer_konzept_gelesen_und_könnte      konzept_gelesen_und_könnte_mir
1                                    1
gelesen_und_könnte_mir_vorstellen    und_könnte_mir_vorstellen_dass
1                                    1
könnte_mir_vorstellen_dass_ich       mir_vorstellen_dass_ich_bei
1                                    1
euch_gut_aufgehoben_wäre_es          ich_hoff_dass_ihr_mir
1                                    1
```

Otra posibilidad de encontrar colocaciones se puede realizar con `textstat_collocations()` del paquete R `quanteda.textstats`. Lo examinaremos con un ejemplo sencillo. Las variables `size` y `min_count` se pueden utilizar para controlar el tamaño de las colocaciones que se van a buscar y su frecuencia mínima de aparición.

```
> # example collocations
> dummytxt <- c(". . . . ich bin schon . .
+ ich bin schon . . . schon deshalb nicht",
+ "ich bin . . ich bin . . ich bin . .
+ ich bin . ich bin",
+ "bin schon deshalb . . bin nicht . bin schon
+ . . . bin schon")
> # standard: size=2, min_count=2
> textstat_collocations(dummytxt)
    collocation count count_nested length lambda z
1   ich bin        7      0          2    5.652 2.746
2   bin schon      5      0          2    4.344 2.532
3   schon deshalb  2      0          2    4.977 2.354
> textstat_collocations(dummytxt, size=3:4)
    collocation count count_nested length lambda     z
1   ich bin schon 2       0          3   -1.11e-16 -3.103e-17
> textstat_collocations(dummytxt, size=2:3, min_count=3)
    collocation count count_nested length lambda z
1   ich bin        7      2          2    5.652 2.746
2   bin schon      5      3          2    4.344 2.532
```

Tarea 10.1: Análisis cuantitativo de textos

Invitamos a los lectores interesados a investigar los análisis anteriores con más detalle en Internet, modificarlos y aplicarlos a un texto más amplio. Para ello sirve el conjunto de datos `data_corpus_inaugural` de `quanteda`, que contiene los discursos de investidura de los de los presidentes estadounidenses.

```
# more and bigger texts
data_corpus_inaugural
summary(data_corpus_inaugural)
head(docvars(data_corpus_inaugural), 10)
```

En el capítulo 5.5.5 hemos visto en el estudio de caso de la educación española que los algoritmos de cluster o el escalado multidimensional en combinación con un análisis de prototipos pueden integrar diferentes aspectos. Aquí no aplicamos estas técnicas. Sin embargo, si tuviéramos muchas cartas de solicitud de este tipo, podríamos examinar a los redactores según la cercanía y la distancia, u ordenar el contenido de las cartas entre sí, clasificarlas, etc. Entonces sería posible comprender qué cartas son más típicas de nuestra clientela y cuáles no. Lo mismo se aplica a las palabras, frases, contenidos, etc. Entonces podríamos examinar en detalle lo que esto significa a nivel cualitativo y, en consecuencia, la entrevista o el programa de terapia a fin de poder responder a estas especiales para poder responder bien a estas particularidades. Se trataría entonces de Métodos Mixtos, incluso combinando teoría y práctica. y práctica.

Sólo trabajamos con un texto. El análisis cuantitativo de textos suele funcionar con muchos textos o con textos bien subdivididos, de modo que se crean unidades de investigación comparables. Pero incluso la aplicación a una unidad de investigación singular muestra aquí que se puede extraer mucho a pesar de contar con cantidades mínimas de datos. Profundizaremos en el contenido más detalladamente en la conclusión de la carta de solicitud. No se debe no negar el análisis cuantitativo de textos desde un punto de vista cualitativo, sino utilizarlo. Nos permite obtener una visión general de grandes cantidades de datos. Incluso con una aplicación concienzuda del paradigma de codificación, no siempre es posible codificarlo todo tan exhaustivamente para poder derivar afirmaciones bien fundamentadas. Así que en lugar de enfrentar un método contra el otro, pensamos más bien en una estrategia "ambos/y" para utilizar los procedimientos en función de su necesidad.

En los análisis de demostración anteriores, debería quedar claro que se puede contar poco con poco texto. Por trivial que pueda parecer, demuestra no obstante que se pueden examinar cuantitativamente casos individuales manejables y limitados, como se muestra en la Figura 10.2. La nube de palabras se refiere a la importancia del egocentrismo en una carta de solicitud de ingreso en un centro de adicciones. La brevedad del texto da lugar a una gran heterogeneidad de los términos, sustantivos y verbos, etc. utilizados. Esto también afecta al paradigma de codificación (véase el capítulo 9.6) un poco, pero no el análisis de secuencias (véase el capítulo 11.13). Así pues, los gráficos de barras sólo muestran cinco términos con una frecuencia superior a 1 (véase la Fig. 10.3). No es casualidad que los dos términos más frecuentes "yo" y "me" apoyen la visión descrita desde la perspectiva del puro recuento. Al fin y al cabo, "tú" viene a continuación, ya que consideramos que los términos "eso" o "estos" en el sentido de stop words (palabras de parada, véase más arriba) son algo menos relevantes para el contenido. Sin embargo, nunca las eliminaríamos, sino que en el contexto de este análisis pospondríamos su examen más detallado, pero nunca dejaríamos de lado una interpretación específica. De este modo, introducimos una priorización del orden según nuestras expectativas de que la referencia en primera persona es más importante que el uso de artículos directos. Sin embargo, no suponemos que los artículos o las palabras de enlace no tengan relevancia.

De este modo, el texto se convierte en el material que queda tras eliminar las palabras de parada. Sobre este material, del que han desaparecido todos los casos individuales, uno puede contar e intentar comprender las frecuencias resultantes y descubrir relaciones entre términos. Lo que se hace entonces es "lo de siempre": buscar y encontrar puntos en común, diferencias, posibles paradojas y contradicciones, así como relaciones entre los términos encontrados y las redes de palabras.

En nuestro ejemplo, algunas de las posibles operaciones textuales enumeradas tienen menos sentido porque el material de datos es demasiado pequeño. Sin embargo, si tomamos los discursos de investidura de los presidentes estadounidenses (véase más arriba) del conjunto de datos `data_corpus_inaugural`, el panorama cambia: En este caso, la tarea podría consistir en reconstruir qué temas y complejos temáticos se encuentran en el centro de los discursos como relevantes. El reto consistiría entonces en llevar las diferencias temporales y el consiguiente cambio potencial de significados y términos asociados, temas, etc. a un nivel comparable para no caer en el cambio temporal de términos. ¿Cómo se presentan los presidentes? ¿Qué es importante para ellos? ¿Cómo ven su propia imagen, su papel, su relación con el Estado y con los ciudadanos? Habría mucho que investigar desde una perspectiva histórica y politológica, por no mencionar la naturaleza del lenguaje y los posibles aspectos psicológicos (pensamiento, sentimiento, motivación y acción) y sociológicos.

Cuidado, sin embargo, con suprimir las stop words supuestamente sin importancia. Ya lo hemos dicho antes, pero nunca se insistirá lo suficiente en ello: Al principio parece tan sencillo que basta con eliminar las palabras de relleno "innecesarias" para luego poder contar las palabras "realmente" relevantes. Con esta mentalidad, se separa el trigo de la paja, porque se conoce las qué del texto que carecen realmente de importancia, si es que existen en absoluto palabras sin importancia. Desde el punto de vista del análisis secuencial (véase cap. 11.2), este tipo de de ponderación de las palabras sería un error mayúsculo, a saber, la falta de igualdad de tratamiento igualitario del material. Y se puede argumentar incluso de forma puramente matemática: La eliminación de palabras, términos, etc. del material total corresponde a una *regularización* (véase. cap. 6.12), en la que, por la razón que sea, se recortan radicalmente y ponen a zero partes de la Prior, por así decirlo. Una consecuencia es que si se elimina esta área de datos a priori ya no se puede interpretar a posteriori. En otras palabras ¡*Los muertos no cuentan cuentos*! En general, los datos que faltan contribuyen muy poco a resolver el problema.

Antes de liberar ciegamente y sin razón toda una lista de palabras para su completa eliminación, hay que pensar seriamente en las consecuencias de esta acción. Ciertamente, hay casos en los que exactamente ese procedimiento está justificado en términos de contenido. Pero habría que poder justificar precisamente este caso. A la inversa, la supresión de las terminaciones y la conservación de las raíces de las palabras puede ayudar a encontrar nuevas relaciones, porque las mismas palabras tienen terminaciones diferentes aquí y allá según la situación. Por supuesto, no queremos pasar por alto algo así. Así que vamos de compromiso en compromiso y posiblemente retrocedemos algunos pasos. Por ejemplo, puede ser que en un caso concreto sea muy significativo un determinado uso de una palabra, concepto, etc. que de otro modo sería irrelevante. Desde una perspectiva de recuento uno puede pasar por alto, pero desde una perspectiva cualitativa es un acontecimiento crítico. Un pequeño y rebuscado ejemplo extraído directamente de nuestra carta de solicitud:

```
[22] "Ich hoffe, dass Ihr mir diese Chance [esta oprtunidad] trotz meinem
relativ hohen Alter"
[23] "gewährt."
```

Ahora decidimos simplemente eliminar todos los artículos como *el, la, los* y los asociados como *este, esto*, etc. El resultado es

```
[22] "Ich hoffe, dass Ihr mir ***** Chance [oportunidad] trotz meinem
relativ hohen Alter"
[23] "gewährt."
```

¿Cuál es la diferencia? Para decirlo brevemente y no con mucho cuidado, simplemente le quitamos la singularidad a la oportunidad abordada y esperada por el cliente potencial. Recordemos que estamos tratando con una persona gravemente adicta que tiene una edad obviamente avanzada (para una persona gravemente adicta) y que ve muy pocas posibilidades de vivir una vida libre del consumo excesivo de drogas y, por lo tanto, de seguir con vida más tiempo. Se trata de una situación cualitativamente diferente a la de los adictos mucho más jóvenes, que posiblemente asumen en su propio exceso de confianza que pueden arreglárselas solos. Este cliente potencial quizás ha pasado por varios intentos de desintoxicación y terapia que no han tenido ningún éxito duradero y no encontramos indicios de un exceso de confianza juvenil. Más bien ocurre lo contrario cuando se piensa en los términos *oportunidad* y mención de la propia vejez. Razón de más para

que *"esta" oportunidad* sea existencialmente relevante. Negarla e ignorar su característica única significa un factor motivador de la terapia y, con él, toda una serie de experiencias personales. Por ejemplo, podríamos preguntarnos por qué y por qué razones el cliente potencial ve *esta oportunidad como "ésta"*. O podríamos preguntarnos si realmente es "*ésta*" en el sentido de "*única*" oportunidad, o si hay otras oportunidades esperando a ser probadas. En este momento no lo sabemos, por lo que no deberíamos eliminar tal dato del registro sin motivo, antes de saberlo con certeza y poder justificarlo en el texto o el caso. De todos modos, sin una revisión concreta del material, es difícil eliminar partes de los datos. Se puede hacer, pero tiene las consecuencias descritas. Y como la referencia a la regulación bayesiana de la distribución previa ha demostrado, esto no sólo tiene razones cualitativas, sino también cuantitativas: El material de datos se modifica y ya no está disponible en su forma original. Si, por el contrario, sólo nos interesan determinados modismos o términos, puede tener sentido concentrarse exclusivamente en ellos. En ese caso, sin embargo, merece más la pena adoptar un enfoque teórico con una lista positiva de palabras relevantes como base de los recuentos. Posteriormente, las palabras clave pueden seguir considerándose en su contexto y éstas, a su vez, contarse según aspectos que deben especificarse (por ejemplo, proximidad y distancia a otras palabras clave). Si se utiliza el enfoque de eliminar partes del material de datos, esta parte desaparece después.

10.1.6 Conclusión de la carta de solicitud

¿Qué hemos sacado en claro de nuestra carta de solicitud? La simple nube de palabras (véase la fig. 10.2) y los diagramas de barras (véanse las figs. 10.1 y 10.3) nos parecen impresionantes, porque ambos muestran que el cliente está preocupado por el "yo", el "mí", el "tú", el "eso" y el "esto". La auto-referencia radical, la perspectiva del ego, se encuentra con el "otro", al que, sin embargo, se tutea y no le trata de usted. Al fin y al cabo, se trata de una carta de solicitud formal y las personas no se conocen. Estamos hablando de una relación laboral profesional y no de un encuentro informal entre amigos. Por lo tanto, un "usted" sería la forma correcta de comunicación y de dirigirse. En consecuencia, el cliente intenta establecer una cercanía que no puede existir, que no existe y que tal vez nunca exista. Al mismo tiempo, se ignora el desequilibrio de poder normal y real entre las personas de la institución y los clientes. Tal vez esto ni siquiera esté en la conciencia, pero sospechamos que simplemente no pasa a primer plano. Todo esto podría apuntar a problemas con las normas y, por tanto, a un posible consumo durante la terapia. Los términos adicionales de "eso" y "estos" podrían interpretarse de forma que, por un lado, la carta tiene estructuras de acción o justificación, es decir, corresponde a un requisito formal de una carta de solicitud, y que varias cosas se nombran con un artículo específico, es decir, contenidos. Sin embargo, como estos contenidos no se mencionan más de una vez debido a la tabla de frecuencias, se trata de términos diferentes. En resumen, esto significa que el cliente potencial da varias razones o puntos de partida por los que debería ser tomado como cliente. Ahora tendríamos que analizar más detenidamente si esto es bueno o malo, por así decirlo. Citar muchas razones podría distraer de lo esencial, es decir, de la motivación básica para la terapia. Una de esas razones podría ser un sincero "Puedo morir pronto si no dejo las drogas". Sin embargo, no se da ninguna razón de este tipo. Por otra parte, esta multiplicidad sugiere recursos cognitivos, porque en la abstinencia en psiquiatría primero hay que inventar algunas buenas razones – a pesar de una experiencia previa ciertamente larga de engañar y camuflar (Gürtler, Studer & Scholz, 2012). La tarea ulterior del análisis consistiría en examinar más de cerca la red de relaciones de estos cinco términos y conectarlos con los demás contenidos. Dado que todos estos otros términos son menciones individuales, es difícil realizar un análisis cuantitativo comparativo. Aquí es donde entran en juego los Métodos Mixtos, para examinar más de cerca estos puntos cualitativamente.

Lo dejamos en este punto con el análisis textual cuantitativo. Hemos basado todas las conclusiones e hipótesis en cinco términos solamente. Hemos podido sacar de ello algo sustancial, incluso más de lo esperado. Quien lo desee puede seguir analizando. Sólo se trataba aquí de mostrar que este enfoque cuantitativo de los datos cualitativos tiene su justificación y que, no obstante, debe utilizarse con cuidado y, desde luego, no de forma semiautomática. Incluso con cinco términos individuales podemos hacer mucho para formar hipótesis.

10.1.7 Ventajas del análisis cuantitativo de textos

Resumamos: ¿qué hacemos realmente en el análisis cuantitativo de textos? Mientras no tengamos una teoría clara y una hipótesis estadística inferencial igualmente clara, el análisis cuantitativo de textos no significa nada más que un análisis exploratorio descriptivo de datos, principalmente de frecuencias de palabras o combinaciones de palabras. Esto no es todavía una teoría, sino simplemente la base de consideraciones teóricas. Sólo cambia el procedimiento cuando ponemos a prueba hipótesis de cualquier tipo basadas en el contenido del texto, de modo que en principio podamos falsarlas o confirmarlas. Pero incluso entonces se aplica lo siguiente: ¡Las hipótesis elaboradas sobre el texto no se comprueban sobre el mismo texto! Deben probarse sobre otro texto. De lo contrario, uno se queda en el terreno de la exploración. Eso en sí no es un problema, porque hay que trabajar con lo que se tiene a mano. Sin embargo, la importancia de las afirmaciones posteriores debe evaluarse de forma realista. Como regla general, dondequiera que se encuentre la palabra "(data) mining", se trata siempre de análisis exploratorios que se deben reproducir en condiciones controladas con hipótesis claras para adquirir alguna importancia. Ni siquiera un enorme conjunto de datos cambia esto en un principio; se necesita una teoría y cuidadosas réplicas. De lo contrario, también pueden surgir falsos positivos y falsos negativos sin ser detectados, ya que siempre hay algo que encontrar en muchos datos en términos puramente estadísticos. Se debe evitar este enfoque escopeta.

Comparemos el procedimiento con el análisis de tablas orientado cuantitativamente del paradigma de codificación. La única diferencia entre ambos enfoques es la unidad de investigación, e incluso eso no está tan claro. Mientras que aquí la atención se centra en palabras, combinaciones de palabras, colocaciones, etc., en el paradigma de codificación se centra en códigos, metacódigos e hipótesis de secuencia, es decir, hipótesis sobre la co-ocurrencia de códigos en cualquier complejidad y abstracción. Todo esto se puede contar, representar gráficamente, producir con y sin texto original, tener en cuenta el contexto, etc. Desde un punto de vista analítico, no hay tanta diferencia. Parece mucho más importante elegir la unidad de investigación adecuada a la pregunta de investigación. La ventaja de las codificaciones es que pueden contener puntos de vista teóricos en el sentido de una interpretación a cualquier nivel complejo, especialmente cuando se comprueba la presencia de secuencias de codificaciones en el texto. Esto parece más difícil con meras palabras y combinaciones de palabras, al menos para las ciencias sociales.

Por el contrario, el análisis cuantitativo de textos trabaja con diccionarios especialmente creados para poder asignarles determinados subconjuntos del material de datos. Este tipo de categorización se corresponde en última instancia con la codificación, sólo que de una forma diferente. Por ejemplo, se podrían crear diccionarios para términos positivos y negativos según el contexto o el contenido con carga emocional, las interacciones sociales, etc. y hacer que se contabilicen. Entonces, en última instancia, volvemos al análisis de la tabla de codificación de una manera indirecta. Esto se podría utilizar para formular hipótesis relacionadas con la secuencia y ponerlas a prueba en el texto. La diferencia no es tan grande, aunque a ambos bandos no les guste oírlo. En cualquiera de los dos casos, paradigma de codificación o análisis cuantitativo del texto, el esfuerzo necesario para llegar a conclusiones sustanciales que estén empíricamente bien fundadas es considerable. Dejaremos de lado los intentos de interpretaciones semiautomáticas y los enfoques de IA, ya que todos ellos no pueden trabajar con inteligencia humana situacional directa.

Pasemos al nivel de la lingüística, que se ocupa menos del significado sociológico o psicológico del lenguaje, sino más por su estructura, sintaxis, etc. El análisis cuantitativo de textos asistido por ordenador parece ser una herramienta muy poderosa. Es posible comprobar estadísticamente si un texto tiene el mismo autor, y mucho más. Las posibilidades aquí son muy amplias, ya que tras una buena preparación del texto se pueden realizar todas las operaciones cuantitativas sobre el texto que estén disponibles. Y eso es realmente mucho.

Capítulo 11

El Paradigma de Reconstrucción

> "By the pricking of my thumbs,
> Something wicked this way comes".

Macbeth, Acto 4º, Escena 2ª, 44-45
William Shakespeare, 1564-1616

11.1. Prólogo

Qué significa para nosotros la cita anterior? ¿Cómo la interpretaríamos y qué cambiaría si no supiéramos que es de Macbeth y de Shakespeare? ¿La interpretaríamos de la misma manera si sólo la leyéramos en alemán y no en el original inglés del siglo XVI? ¿Cómo lo traduciríamos? ¿Cómo influyen estas opciones en la interpretación del contenido? ¿Hay alguna diferencia en la ¿Hay alguna diferencia en la interpretación del contenido manifiesto frente al latente? El paradigma de la reconstrucción ofrece una respuesta clara a algunas de estas preguntas. contenido latente-el paradigma de la reconstrucción ofrece una respuesta clara e inequívoca. Este se proporciona en el curso de una interpretación secuencial del contenido latente y sobre la base de una pregunta. El paradigma de reconstrucción no tiene respuesta para otras cuestiones, como la de la traducción. Pero podríamos interpretar distintas traducciones por separado y comparar los resultados. Eso sería muy emocionante, porque nadie ha llevado a cabo una investigación de este tipo. La cuestión de la traducción tiene sin duda una influencia relevante - Frank Günther (Shakespeare, 2014) traduce las líneas anteriores de la siguiente manera

> "Por el picor del pulgar,
> aparece alguien malvado".

A pesar de ello, bien podríamos traducir nosotros mismos la cita como

> "Por la picazón de mi pulgar,
> Algo malo viene hacia aquí".

¿Cuál es la posible diferencia en el significado latente del texto? En el paradigma de la reconstrucción, la pregunta es "¿De qué trata realmente el texto?". Así pues, podemos preguntarnos: ¿estamos reconstruyendo el contenido latente de "Por el picor del pulgar" de la misma manera que "Por la picazón de mi pulgar"? ¿En qué contextos utilizaríamos una expresión, en qué otras situaciones utilizaríamos la otra; y qué cambia por el diferente significado en cada caso? ¿Cuál de estas interpretaciones reconstruibles corresponde al original inglés? Esto es importante porque en el paradigma de la reconstrucción nos ceñimos a lo literal, es decir, a lo que realmente se dice y no a lo que nosotros creemos que significa o debería significar. Una estricta orientación hacia la realidad guía el análisis.

Para ello necesitamos una sólida comprensión del lenguaje y la cultura, así como disciplina en el análisis guiado por reglas. Si no queremos ver Macbeth cómodamente en el teatro, sino reconstruir su contenido

meticulosamente, nos concentramos en el texto tal como es. Al hacerlo, se nota que en el mundo germanoparlante – a no ser que también seamos angloparlantes nativos – tenemos un conocimiento intuitivo de nuestra lengua y del significado de las palabras. Sin embargo, no tenemos necesariamente esta comprensión y, desde luego, no a este nivel para la lengua inglesa. Mientras no tengamos esta comprensión intuitiva de la lengua, seremos por tanto cautos a la hora de intentar interpretar estas líneas del inglés antiguo de forma inocente. Una breve interpretación antes del contexto del Acto IV puede encontrarse en un blogpost de Karae Jacobi (2015). Citamos aquí deliberadamente un sitio web de análisis literarios de trabajos trimestrales y proyectos de investigación de estudiantes, ya que ilustra que un enfoque basado en la reconstrucción es bastante diferente. Así, en términos literarios la escena con el trasfondo del inglés antiguo, la acción del acto IV (Macbeth ya se ha convertido en asesino y traidor y está a punto de hacer cualquier cosa para conservar la corona) y el uso común de este modismo (a saber, "tener un mal presentimiento"). Precisamente la asociación de sensaciones dactilares y mal presagio fue sin duda comprendida por todos los públicos de la época de Shakespeare. Sin embargo, si utilizamos esta información para la interpretación, reducimos drásticamente nuestro abanico de posibles interpretaciones y, en consecuencia, tendemos a buscar menos alternativas reales. Pero queremos utilizar toda la información por igual y luego decidir cuál es la interpretación que mejor puede explicar la información. La forma de utilizar, así como la secuencia y el orden de la información disponible (léase: el texto como tal) hacen del paradigma de la reconstrucción no sólo una excelente metodología para interpretar textos y cualquier información en general, sino también un procedimiento científico muy preciso.

Así, podríamos considerar que el mal significado es o bien un espíritu y, por tanto, algo abstractamente numinoso, o en la forma de un animal o de un ser humano, o si el enunciado no está destinado a ser real en absoluto, o si se trata de un meta-comentario o incluso de un enunciado auto-referencial. autorreferencial. ¿Quizá Macbeth está filosofando sobre sí mismo? El camino ("this way comes") puede ser una forma de vida, la dirección elegida de un proceso o de un marco general (de algo que actualmente desconocemos), etc. Esto ya da lugar a algunas combinaciones de interpretaciones. Los lectores interesados pueden formular estas hipótesis en este momento escribirlas y cotejarlas con el texto de Macbeth en su conjunto. Pongamos a prueba estas hipótesis echando un vistazo al siguiente pasaje y veamos cuáles de las hipótesis siguen teniendo sentido a la vista del nuevo pasaje y cuáles deberíamos cambiar. La siguiente frase del texto dice

> „Open, locks,
> Whoever knocks."

Por tanto, probablemente no se trate de animales ni de cosas numinosas, porque rara vez llaman a la puerta, salvo en los cuentos de hadas (el lobo malvado...). Así pues, es más probable que el mal se asocie con un ser o alguien que puede llamar a la puerta. Cuando alguien llama a la puerta, es probable que el "camino" del pasaje anterior sea literal y no metafórico. Así pues, alguien se acerca realmente. Pero, ¿se trata de una persona malvada o del mal en forma humana? Ahora podríamos y deberíamos preguntarnos además, ¿qué personas visitan a otras personas suponiendo que son realmente malvadas? Por tanto, es de suponer que el hablante tiene un profundo conocimiento del propio mal, [...] - que sea tarea del lector formular aquí otras hipótesis de la manera descrita, para ir desentrañando el sentido latente de la acción y cotejarlo con pasajes posteriores del texto.

Como se puede ver, el paradigma de la reconstrucción procede de un modo completamente distinto al de una interpretación literaria convencional o en el paradigma de codificación (véase el capítulo 9), en el que la información posterior o la información contextual se utiliza específicamente para formar categorías y suposiciones sobre los contextos. Aquí, en cambio, ignoramos inicialmente el contexto y no utilizamos conocimientos previos sobre el estado mental en el que ya se encuentra Macbeth y qué ideas malignas está formando en su interior. Ignoramos deliberadamente el hecho de que el modismo que estamos examinando ha pasado a formar parte del conocimiento común de los malos augurios. Así, según las ideas supersticiosas los dolores o sensaciones extrañas sin causa directa se consideraban indicios sobrenaturales de la llegada de acontecimientos venideros. Pretendemos deliberadamente ser "ingenuos" e "ignorantes" para no fiarnos de nuestras suposiciones implícitas e intentos inconscientes de confirmación, sino para sucesivamente interpretar el texto en su orden secuencial natural. Sin embargo, *no más tontos* de lo necesario. Más bien, nos limitamos a estructurar la información de otra manera, a saber, que se interprete *tal y como surgió de forma natural*.

En el caso de Macbeth, esto significa interpretar primero la declaración de la bruja antes de analizar después el aspecto y las expresiones del propio Macbeth, y no, a la inversa, utilizar acontecimientos temporalmente posteriores para analizar acciones y declaraciones temporalmente anteriores.

También omitiríamos el hecho de que "algo malvado" se refiere al propio Macbeth, lo cual es un giro interesante dado que una bruja hace esta declaración. Al fin y al cabo, las propias brujas se asocian a menudo con "malvado", pero no siempre, por supuesto. No – cuando trabajamos según el paradigma de la reconstrucción, empezamos a interpretar irreflexivamente a *partir de la unidad textual que tenemos*. No utilizamos información de otras partes del texto, sino que tomamos lo *que es el caso* y trabajamos a lo largo de cómo el caso se desarrolla y desenvuelve naturalmente. No nos basamos en absoluto en interpretaciones generales de libros conocidos, sino que relacionamos cada palabra y cada frase de tal manera que podamos formar nos hipótesis sobre de qué se trata en realidad. Y si no podemos formular más hipótesis porque se ha agotado el actual espacio de posibilidades de interpretaciones potenciales, sólo entonces nos fijamos en la siguiente unidad textual y examinamos críticamente si nuestras hipótesis siguen siendo válidas a la vista de los nuevos pasajes del texto y cuáles de ellas. Nos preguntamos qué hipótesis debemos abandonar, cuáles deben modificarse, etc. Y repetimos este juego hasta que no surjan nuevas interpretaciones sustanciales. Estructuralmente, esto es muy parecido a la saturación teórica de la teoría fundamentada (Glaser y Strauss, 1967). Normalmente nos encontramos entonces en la fase de poder formar una *hipótesis preliminar viable sobre la estructura del caso*. Y a continuación intentaríamos hacerla recaer específicamente en otros pasajes del texto. La selección de este pasaje del texto ya no está ligada a la secuencia, es decir, también podemos tomar un acto anterior de Macbeth para la falsación, siempre y cuando aún no se haya incluido en el análisis. Este proceso interpretativo no es literario en absoluto. Es más bien un proceso evolutivo y nos recuerda la teoría de la evolución de Darwin: las hipótesis compiten y se desarrollan y cambian a lo largo del texto. Pero no hay ni meta ni planificación, ni nadie que prescriba lo que el contenido latente y, por tanto, el producto del proceso evolutivo. Somos *responsables del proceso* y nos lo tomamos muy en serio.

Resumamos el punto de partida: el paradigma de la reconstrucción nos permite reconstruir la estructura subyacente del texto sobre la base de la falsificación paso a paso, la aplicación de la teoría evolutiva darwiniana y siguiendo estrictamente el orden natural del texto. Esto incluye tanto la fase de e*xploración*, la fase de *comprobación de hipótesis* de modo de falsificación evolutiva y la fase de *confirmación* concreta de la estructura encontrada en pasajes del texto que aún no han sido examinados. El análisis concreto procede de la sociología y se denomina *análisis de secuencia* (Oevermann, 1993, 2000; Oevermann, Allert & Konau, 1980; Wernet, 2000). Una gran parte del texto siguiente trata de ello. El análisis de secuencia, a su vez, es el núcleo central de la *metodología de la Hermenéutica Objetiva* (Oevermann, Allert, Konau & Krambeck, 1979a). En su versión original, la Hermenéutica Objetiva es tanto una *teoría sobre el significado estructurado del mundo* como una *metodología de investigación para la reconstrucción del sentido del mundo social*, sus elementos y las múltiples relaciones entre ellos.

11.2 Hermenéutica objetiva y análisis de secuencias

El fundamento teórico del análisis de secuencia reside en la metodología de investigación de la Hermenéutica Objetiva según Ulrich Oevermann y colegas. Este enfoque intenta abordar cuestiones fundamentales de la investigación social, así como utilizar los conocimientos, resultados y métodos del análisis de textos en campos de trabajo prácticos. La hermenéutica objetiva no es sólo un método para analizar protocolos (textuales), sino una estrategia de investigación en sí misma, con el fin de llegar a resultados significativos desde el punto de vista socio-científico, es decir, garantizados intersubjetivamente. El objetivo *no es determinar el significado subjetivo* de las acciones (de interacción) y sus consecuencias, de las que una persona es consciente y que puede rectificar, sino más bien *reconstruir el significado objetivo* de estos actos expresivos. Una gestalt expresiva es la manifestación real de las disposiciones subjetivas. A ello precede la práctica de la vida, en el curso de la cual se manifiestan las respectivas formas de expresión. Oevermann (1996b, p.2) subraya,

> "que toda disposición subjetiva, es decir, todo motivo psicológico, toda expectativa, toda opinión, actitud, orientación de valores, cada imaginación, afinación, fantasía y cada deseo nunca es metódicamente verificable directamente, sino siempre sólo mediante una forma expresiva o una huella en la que se encarna o que ha dejado tras de sí".

Esto refleja una actitud conductista y orientada a la acción habitual en las ciencias sociales, a saber, que las cosas relevantes (motivos, sentimientos, pensamiento) no son directamente observables, sino que deben inferirse a partir de indicadores observables. Así pues, la Hermenéutica Objetiva pertenece al conductismo social.

11.3 Práctica y práctica vital

El término *protocolo* es aplicado por Oevermann (2000) a todos los datos registrables y a sus formatos. Las unidades de investigación de la Hermenéutica Objetiva se denominan, por tanto, como *protocolos de la realidad*. Son formas expresivas de los datos derivados de razones teóricas o prácticas de la estructura deel caso, en que los investigadores se interesan. Los protocolos pueden ser de origenes muy diversos, como documentos escritos, imágenes, secuencias de movimiento, secuencias de sonido y tono o interacciones sociales. Oevermann, Allert, Konau y Krambeck (1979a, p.378) dicen:

> "El objeto concreto de los procedimientos de la hermenéutica objetiva son protocolos de acciones o interacciones sociales reales, simbólicamente mediadas, ya sean escritas acústicas, visuales, combinadas en diferentes medios o fijaciones archivadas de otro modo".

Afortunadamente, este punto de vista no ha cambiado realmente con el paso de los años. Así escribe Oevermann (2002, p.1) casi 25 años después:

> "El objeto central de la metodología de la hermenéutica objetiva son las estructuras de sentido latentes y estructuras objetivas de sentido de las figuras expresivas, en las que se presentan a nosotros como científicos empíricos del mundo estructurado según los sentidos solamente los fenómenos psicológicos, sociales y culturales, y en las que nos encarnamos a nosotros mismos, así como representamos el mundo de la experiencia que tenemos enfrente".

El objetivo no es la realización de "un sentido subjetivo que pueda ser comprendido, sino más bien ... [se trata del] sentido objetivo generado por reglas que caracteriza las acciones prácticas y sus objetivaciones" (Oevermann, 1993, p.251). La forma expresiva (protocolo) y la práctica vital (realidad protocolizada) están dialécticamente conectadas entre sí. Oevermann (1997, p.7f.) comenta sobre la práctica vital,

> "La práctica vital se ve así permanentemente confrontada con decisiones que se debe justificar racionalmente. Las decisiones auténticas, sin embargo, son sólo aquellas en las que la elección racional no está ya fijada en el momento de la decisión como en un cálculo racional. Más bien se tiene tomar decisiones hacía un futuro abierto, y su justificabilidad racional debe demostrarse de hecho sólo en ese futuro abierto".

En la Hermenéutica Objetiva, el término *práctica*, que se utiliza con frecuencia, se refiere a la totalidad de la estructuración del sentido en los espacios vitales humanos (acciones) y se muestra en las figuras de expresión conccretas, las manifestaciones reales como *práctica vital concreta*. Las formas expresivas siguen viviendo casi siempre en espacio y tiempo en un nivel abstracto a través de su reconstrucción mediante el análisis de protocolos (textos) – y son accesibles para un examen posterior en cualquier momento. Mientras que la conciencia individual haya olvidado hace tiempo estas figuras expresivas, su reconstrucción permanece para la posteridad en forma de huellas (Studer, 1998, p.27, fig. 2.2.2-1). La práctica vital como parte complementaria de la praxis se reduce a las entidades sociales (ibíd., p.33),

> "que, en virtud de su estructuración, tienen la propiedad de generar autonomía y mantener el centro constructor de la acción ... La práctica vital se erige aquí como una cosa abstracta en la que se concibe el estrato uniforme de la acción práctica que funda todas las expresiones sociales de la vida, que al mismo tiempo también constituye la base de una interpretación materialista de la historia del género, de la historia y de la epistemología".

La *práctica vital* tiene una fuerte referencia histórica, que según Oevermann (1997, p.7f.) se mueve permanentemente entre dos polos:

> "La práctica vital se constituye ahora como autónoma precisamente en esta unidad contradictoria de la compulsión decisoria y la obligación de la justificación. Sólo se manifiesta como tal en la crisis".

La práctica de la vida apunta psicológicamente a un nivel disposicional habitual de comportamiento complejo y acción en el contexto del vínculo inseparable entre "tener que actuar" (= compulsión a decidir) y "tener que justificar la acción" (= obligación a justificar). La integración en el contexto de la práctica concreta de la vida anclada biográficamente tiene lugar a lo largo de la dialéctica según Hegel. Si esta figura potencialmente paradójica fracasa sin integración, es casi seguro que tarde o temprano se producirá una crisis. Es importante comprender por qué la crisis se produce en tal o cual momento de la biografía individual:

Recordatorio 11.1: Práctica vital y crisis

La práctica vital tiene éxito mientras los hábitos sean suficientes. Si los hábitos ya no son suficientes debido a cambios internos y externos, los hábitos deben cambiar, lo que a su vez requiere la integración de figuras contradictorias. Si esta integración fracasa, se produce la crisis.

11.4 Reconstrucción a partir de las huellas

El sentido objetivo comprende, pues, el ámbito dentro del cual lo que se quiere decir subjetivamente ("eso es lo que quiero...", "eso es lo que quiero decir...", etc.) se expresa y deja sus *huellas en la realidad*. Estas huellas corresponden a las acciones e interacciones reales en la realidad. En la Hermenéutica Objetiva, el concepto de *objetividad* implica que, mediante la aplicación de operaciones metódicas una realidad es tan claramente demostrable que esta interpretación puede calificarse de objetiva. La objetividad no es un requisito abstracto casi "deshumanizado" como se suele favorecerlo en la investigación experimental, sino más bien un resultado de la aplicación cuidadosa y reproducible de la metodología de investigación con el correspondiente trabajo de justificación.

El proceso hacia ello es de naturaleza intersubjetiva, idealmente a través del análisis en el grupo. Analizado sólo se pueden analizar las huellas de la práctica vital, no la propia práctica tal y como se manifiesta de momento en momento. Esto, sin embargo, afirma una pretensión de realidad del análisis que, por lo demás, sólo reclaman para sí las ciencias naturales. No todos los autores comparten este punto de vista con tanta radicalidad. Existen posturas contrarias (Groeben, 1986, p.157s.). Éstas subrayan que no puede haber objetividad en el análisis cualitativo, sino que el concepto de intersubjetividad como negociación de la construcción de significados y la triangulación (Flick, 2000) como estrategia de cambio de perspectivas proporcionan una imagen más realista de lo que está ocurriendo. Cuestionable si esto minimiza la pretensión de validez, ya que sostenemos que no hay más que la intersubjetividad fundada en la teoría y el empirismo. Eso es un problema sólo si no se contrasta críticamente la (inter)subjetividad con la realidad.

Groeben (1986) llama al método de Oevermann una hermenéutica monológica, que él contrasta con una hermenéutica dialógica (Scheele et al., 1988; Scheele & Groeben, 1988) para integrar el monismo y el dualismo. Se llega a un compromiso que *subordina* el análisis cualitativo a un posterior análisis cuantitativo "suave". La crítica de esta postura es que dicho procedimiento no utiliza el potencial original del análisis cualitativo de datos, ya que se le asocia a la pura subjetividad. Por otra parte, tampoco se explotan plenamente las posibilidades del "análisis estadístico duro de datos". La Hermenéutica Objetiva no requiere el análisis de datos estadísticos para abarcar todos los aspectos de la epistemología (véase el capítulo 1.2) y, al mismo tiempo, evita el ámbito de la pura subjetividad y las "opiniones". Debido a la intersubjetividad anclada en el proceso y el enfoque de falsificación, que trabaja estrictamente a lo largo de los hechos verificables, el proceso de análisis conduce a una hipótesis de estructura de casos con base empírica. Desde el punto de vista del constructivismo o del postmodernismo, la objetividad no es un concepto fiable ni siquiera realizable. Por el contrario, se hace hincapié en el trabajo colaborativo en los procesos de formar conceptos y en el lenguaje. Esto tiene consecuencias para la interpretación de los datos (textos) y la disolución de la forma, que llama la atención sobre la incrustación sociocultural de la ciencia. Desde nuestro punto de vista, éste es una cosa extremadamente relevante a tener en cuenta en la interpretación y aplicación de los análisis objetivo-hermenéuticos. En un mundo de verdad relativa, no sólo cambian las perspectivas de las personas a lo largo del espacio y del tiempo, sino también los niveles de significado a un nivel más abstracto – lengua, cultura y sociedad. No tenemos forma de "trascender" simplemente estas limitaciones. Como mucho, podemos reformularlas y modelarlas como incertidumbre. Esto queda claro cuando volvemos a la cita inicial de Macbeth ("By the pricking of my thumbs, Something wicked this way comes.") e interpretarla, por un lado basado en nuestra *moderna* comprensión del lenguaje, y por otro (si pudiéramos darnos cuenta) basado en la comprensión del lenguaje del siglo XVI. Nosotros estamos seguros de que las interpretaciones difieren sustancialmente.

Volvamos de nuevo a la discusión sobre la objetividad. Rechazando los argumentos de que los análisis objetivo-hermenéuticos no merecen el calificativo de "objetivos" se podría rebatir que a través de la forma rigurosa y probatoria del análisis, la intersubjetividad alcanza un grado tan alto que el término objetividad es legítimo, ya que aquí se cruza una frontera cualitativa. Más no es posible en las ciencias naturales, donde la objetividad significa una secuencia de convenciones de interpretaciones intersubjetivas de la teoría y el empirismo. Se podría preguntar ahora qué pasa con la IA, si aquí surge un elemento subjetivo, aunque esté puramente basada en datos. Los algoritmos subyacentes siguen estando hechos por el hombre. No se trata, por tanto, de un acuerdo de algún modo sólo condicional de los científicos sobre una estructura de casos, sino *regida por reglas y reproducible de un procedimiento cuyos pasos están ahora completamente controlados* por la gestión de datos y documentación por ordenador (s. cap. 11.12 o A.), que se puede ahora reproducir completamente. En principio, la misma "objetividad" y transparencia que también puede se lograr en el marco de análisis estadísticos u otros entornos de investigación estrictos (por ejemplo, experimentos). El argumento de que el sujeto de la investigación, la persona estudiada, no tiene voz ni voto aquí es correcto. Pero como la Hermenéutica Objetiva no pretende decir algo sobre el mundo interior subjetivo de los sujetos de la investigación, ni de poder decir algo en absoluto, esta objeción parece quedar en nada. No es de esto de lo que trata la Hermenéutica Objetiva. No se trata de lo que ocurre en las personas y de cómo se ven a sí mismas cuando actúan, sino de lo que causa la conjunto de factores que se pueden reconstruir desde el exterior cuando las personas actúan. A partir de ahí, se pueden hacer predicciones y, en principio, ponerlas a prueba en acciones futuras. Por desgracia, esto no se practica sistemáticamente – si es que se practica – en el campo científico de la Hermenéutica Objetiva, pero esto no es un obstáculo. Común es la incrustación en contextos reales de acción, por ejemplo en la práctica terapéutica, cuando una estructura de caso reconstruida se convierte en el punto de partida para planificar intervenciones complejas y las derivaciones de la misma enmarcan y determinan adaptativamente el proceso terapéutico. O cuando el resultado de la carta de solicitud redactada a partir de la abstinencia en psiquiatría para una terapia (de adicción) es ya tan revelador sobre quién acudirá y qué recursos, problemas y retos son de esperar (Studer, 1995; Gürtler, Studer & Scholz, 2012). Como resultado, sin embargo, tal interpretación debe adaptarse siempre a los cambios del proceso, lo que a su vez afecta a las interpretaciones posteriores (por ejemplo, de la biografía, el genograma, ...) y, por tanto, en la propia estructura del caso en sí (Studer, 1998). De lo contrario, la terapia no cumpliría su pretensión de iniciación al cambio.

Incluso Groeben (1986) estableció específicamente la distinción entre comportamiento, hacer y acción, para poder clasificar la validez incierta de las afirmaciones subjetivas. Esta distinción tiene en cuenta el hecho de que las personas pueden a veces, pero no siempre, decir algo sobre sus verdaderas razones para actuar (motivos, intenciones). A veces lo que la gente dice guía sus acciones, y a veces no es así, y muchas veces la verdad se encuentra entre medias. Por lo tanto, nunca basta con creer únicamente en la declaración subjetiva de una persona, sino que ésta es una variable dinámica, cuya validez se debe investigar adaptada a la situación. Esto no significa no respetar las afirmaciones como subjetivamente serias. Significa simplemente cuestionar si las intenciones y motivos expresados son realmente los efectivos cuando las personas actúan. Si es así, las motivos son expresiones legítimas de interés, pero no efectivas en términos de acción. En última instancia, esto conduce a la dicotomía en el proceso de investigación de Groeben, ya que carece de la metodología para comprobar la validez real de las expresiones subjetivas, para lo cual utiliza en una segunda fase métodos estadísticos "suaves" con el fin de superar la estrecha estructura de una lógica estadística experimental y estrictamente inferencial. Desde el programa de investigación "Teorías Subjetivas", que él fundó, le concibió específicamente como contrapartida del conductismo y por lo tanto no surgió de una posición libre, lo que naturalmente siempre tiene un efecto constrictivo, se niega naturalmente un planteamiento conductista social, tal como se practica en la Hermenéutica Objetiva. Groeben intenta eludir esto mediante una hermenéutica dialógica acercándose primero a las personas en pie de igualdad en principio para a fin de reconstruir junto con ellas sus teorías subjetivas. En un segundo paso, se examina esta teoría subjetiva para orientar la acción, siguiendo una lógica procedimental estadística más suave. En cambio, la Hermenéutica Objetiva ya realiza este segundo paso en el primero, por lo que en el primer paso los sujetos de la investigación no tienen nada que decir. Metodológicamente, sin embargo, no hay ningún obstáculo para conectar la visión subjetiva y la intersubjetiva – objetiva – reconstrucción entre sí. De nuevo – no se necesitan estadísticas ni siquiera entonces, sino una pregunta de investigación formulada adecuadamente.

Si damos un paso atrás, la Hermenéutica Objetiva muestra similitudes estructurales con el conductismo y la teoría conductista, ya que está estrictamente orientada a las huellas realmente observables de la acción humana. Sin embargo, a diferencia del conductismo el comportamiento no se documenta y comprende simplemente sobre la base del condicionamiento, sino basado en las expresiones vitales, los protocolos de la realidad con todas las expresiones de los sujetos investigados. Este protocolo (textual) se convierte entonces en el objeto del análisis. El marco metodológico se mantiene estricto para reconstruir el significado latente dentro de los protocolos. Todo esto no tiene lugar en el conductismo, ya que éste carece precisamente de esa metodología de investigación para el análisis de datos.

Se puede encontrar una justificación detallada del complejo marco teórico de referencia, entre otros lugares, en Oevermann (por ejemplo, 1979b, 1981, 1993, 1996a) y Oevermann, Allert, Konau y Krambeck (1979a). Estos trabajos proceden en su mayoría de la sociología (Oevermann, 1979b, 1981, 2002), la pedagogía (Oevermann, 1996b), el trabajo clínico (Oevermann, 1998a, 2000) y la criminología (por ejemplo, Reichertz, 2002). La Hermenéutica Objetiva es en realidad inexistente en la psicología y es prácticamente inexistente fuera de la zona de habla alemana. A largo plazo, una situación tan desfavorable conduce a la desaparición de un enfoque, independientemente de su calidad.

11.5 El análisis de secuencia como herramienta central de investigación

> "El análisis de secuencias se acurruca en la estructura básica de los acontecimientos humano-sociales reales y, por eso, no es un método externo al objeto, sino un método que corresponda a la materia misma y es adecuada a ella a diferencia de los métodos habituales de medición y clasificación. De hecho, en la vida práctica se tiene que decidir, en principio, en cada punto de la secuencia entre las opciones aún abiertas para el futuro. Sin embargo, en la gran de los casos esto ocurre de forma subjetiva, como si pasara desapercibido, sobre la base de rutinas que originalmente eran soluciones de crisis" (Oevermann, 2002, p.33).

11.5.1 Análisis secuencial y estructura del caso

La metodología más destacada de la Hermenéutica Objetiva es el procedimiento analítico del análisis de secuencias. "Se apoya en la secuencialidad que es constitutiva de la acción humana" (ibíd., p.6). Sin embargo, secuencial no significa simplemente trabajar de adelante hacia atrás. Se trata más bien de seguir estrictamente la secuencia del texto a analizar para abrir nuevas posibilidades de interpretación y volver a cerrarlas cuando no resistan el escrutinio del texto. El término "lectura" se utiliza en la Hermenéutica Objetiva para designar la multitud de posibilidades de interpretación. El análisis secuencial es, por tanto, la interacción entre la posibilidad y la realidad y examina si y cuáles de las posibles lecturas se pueden encontrar en el texto y, en caso contrario, cómo se deben modificar las lecturas para que sean coherentes con el texto. El orden secuencial del análisis resulta de la estructura genérica del lenguaje, de la que tenemos una comprensión intuitiva, al menos para nuestra lengua materna. La estructura de casos propiamente dicha es el tipo de texto analizado que representa la estructura latente y puede expresarse con palabras. Mientras esta estructura de casos no esté realmente asegurada empíricamente y comprobada críticamente, se denomina hipótesis de estructura de casos.

11.5.2 Control metodológico y falsación

Dado que el mundo está en constante cambio, los procesos o acciones concretos nunca están disponibles para su comparación crítica con un desarrollo alternativo en un universo paralelo. No existe la posibilidad de un análisis simultáneo y paralelo "en vivo". Esto lleva a la necesidad de concentrarse en la investigación de protocolos, es decir, textos, de forma metodológicamente controlada. A la inversa, esto abre la posibilidad de poder examinar intersubjetivamente estructuras de sentido ya reconstruidas sobre el mismo u otros protocolos. En este sentido, la Hermenéutica Objetiva reclama en principio para sus conclusiones la misma forma de objetividad a la que nos tienen acostumbrados las ciencias naturales. Esto se apoya en la exigencia de una capacidad de principio para fallar en la mediación entre la realidad concreta y la posibilidad abstracta (Studer, 1998). Esta exigencia es doblemente válida – tanto como cualidad metodológica para el propio proceso de análisis como igualmente desde la teoría (del objeto) de la Hermenéutica Objetiva con respecto a la práctica vital humana concreta y su afrontamiento sobre todo en crisis. Este espacio de posibilidad para la práctica vital se vuelve problemático cuando deja de existir y cuando dejan de existir los diseños vitales. La unidad contradictoria de la presión de decidir y la obligación de justificar (véase más arriba) se manifiesta entonces en la crisis, puesto que los hábitos y las rutinas ya no funcionan. La existencia del espacio de posibilidad siempre implica que "todavía podría ser algo que no era" y "podría actuarse de una manera que todavía no se ha actuado".

Mientras que en la práctica concreta de la vida el ser humano se ve obligado a crear algo nuevo y a superar las propias rutinas para superar la crisis, esta figura básica se aplica igualmente al propio proceso de análisis: Las interpretaciones no son rutinas o hábitos, sino el espacio de tensión entre la realidad (texto) y la posibilidad (hipótesis) que conduce idealmente a la crisis para desactivar de ella los propios presupuestos y reconstruir lo que realmente contiene el texto. Esto requiere un control metodológico y una comprobación crítica de la realidad, es decir, una falsificación. Así, las hipótesis se examinan críticamente sobre el texto, se mantienen, se modifican o se rechazan. Al tratar las hipótesis, Oevermann sigue estrictamente un procedimiento de falsación que Popper (1943) formuló en el racionalismo crítico. En este sentido, la Hermenéutica Objetiva gira sobre sí misma, como debe ser.

En concreto, un planteamiento falsificatorio significa abrir toda la posibilidad de reconstruir o interpretar el sentido y confrontarlo con el texto o con una parte negada del texto. Ante este examen crítico del texto, la mayoría de las hipótesis que abren el espacio de posibilidades se derrumban. Por lo tanto, se falsifican y se descartan. Otras necesitan revisión y reformulación o incluso se demuestran temporalmente sin más cambios y se mantienen sin cambios. Este proceso se repite hasta que sólo una hipótesis viable y ahora complejamente formulada (estructura de caso) ha prevalecido evolutivamente sobre todas las demás hipótesis

del espacio de posibilidades. Sin embargo, a diferencia de variables como los factores de Bayes (véase el capítulo 6.8.1), las hipótesis de estructura de casos contienen un alto grado de complejidad, ya que también incluyen casos especiales y variantes que desempeñan un papel en el texto. No son un modelo que sea sólo relativamente mejor que otro, sino un modelo complejo que puede explicar el texto de forma exhaustiva. De este modo, cumplen los requisitos que el estadístico Andrew Gelman impone a los modelos.

A continuación se realizan pruebas de falsación específicas sobre partes del texto que aún no se han examinado, lo que equivale aproximadamente a predecir modelos sobre nuevos datos en estadística. Podemos imaginar esto con la imagen de un cocodrilo que primero abre su enorme boca al máximo y luego la deja colapsar con toda su fuerza. El cocodrilo repite esto hasta que toda la sustancia utilizable ha sido aplastada y depositada en la boca y digerida, es decir, eliminada. Lo que queda es la estructura de nuestro caso, que podía y puede resistir los ataques selectivos y repetidos de la boca del cocodrilo.

11.5.3 El significado latente

El *análisis de secuencia* es la herramienta de investigación elegida para llevar a cabo reconstrucciones precisas de casos y derivar interpretaciones y conclusiones adecuadas al contexto (Wernet, 2000). (Wernet, 2000). Lo que hace especial al análisis secuencial es que, como su nombre indica, es un *enfoque secuencial* del análisis. Se lleva a cabo siguiendo la *estructura natural* del texto estudiado, es decir, tal y como surgió *realmente* en su orden cronológico. Esta búsqueda de estructuras latentes significativas es, por tanto, la meta central de la Hermenéutica Objetiva. Está estrechamente relacionada con los términos *significado objetivo*, *significado latente* y *disposición subjetiva*, que se deben reconstruir de forma intersubjetivamente validada e inequívoca y de los que ya hemos hablado.

El adjetivo *latente* se refiere a su vez al hecho de que las estructuras reconstruidas *no están sujetas a la conciencia* y no son en absoluto accesibles reflexivamente a las personas implicadas. Una vez reconstruidas las estructuras objetivas se puede analizar el significado subjetivo de las acciones con mucho más detalle.

La reconstrucción de las estructuras objetivo-latentes de significado es posible gracias a la suposición de que un lengua tiene reglas y significados compartidos intersubjetivamente. En un nivel lógico-analítico de reconstrucción, las estructuras latentes de significado que se deben reconstruir son independientes de la realización manifiesta (real) de estas estructuras supuestas en la estructuras en la conciencia de los agentes. Estas realidades latentes son abstractas no perceptibles directamente por los sentidos, pero verificables como experiencia con ayuda de reglas metodológicas explícitas. Debido a estas reglas genéricas y generalmente válidas del lenguaje, Oevermann llama *objetivas* a las estructuras sensoriales reconstruidas. Por el contrario, rechaza cualquier afirmaciones y suposiciones sobre el mundo subjetivo de la experiencia de los investigados mismos. Considera que éstas no pueden ser reconstruidas por investigadores externos. Esta postura de la incapacidad respecto a los procesos internos es compartida por la Hermenéutica Objetiva y el conductismo psicológico sin ser, no obstante, idéntica a ella. La Hermenéutica Objetiva así es muy cerca al conductismo social según George Herbert Mead (1863-1931).

11.6 El ámbito de la Hermenéutica Objetiva

La Hermenéutica Objetiva es un paradigma de investigación por derecho propio, cuya comprensión va mucho más allá del análisis de secuencia. Se ocupa de

- la práctica humana y la práctica vital,
- la relación entre teoría y práctica,
- la mediación de lo que sólo puede hacerse en la práctica (Oevermann, 1998b) así como

- las teorías sociológicas de constitución y socialización, sobre las que Oevermann y sus colegas han publicado ampliamente (por ejemplo, Oevermann, 1996b, 2004).

En estrecha relación con esto se encuentran conceptos dialécticos como crisis y rutina o la necesidad de tomar decisiones y la necesidad de dar razones (véase más arriba). Todo ello sirve para elaborar adecuadamente la práctica humana y la práctica vital en vista de la fugacidad del mundo junto con los contenidos de sentido latentes subyacentes.

En lo que sigue no profundizaremos en estas complejísimas definiciones teóricas de términos, justificaciones y conexiones dentro del discurso de la ciencia. Tampoco nos extenderemos más en la teoría de Oevermann sobre los rasgos constitutivos de nuestro mundo estructurado por los sentidos. Se pueden encontrar explicaciones más detalladas de la Hermenéutica Objetiva en, entre otros, Oevermann, Allert, Konau y Krambeck (1979a) u Oevermann (2000) y muchas otras publicaciones y manuscritos del autor y sus colegas. Sin embargo, están formulados de forma decididamente abstracta-científica y su contenido está orientado sobre todo a la sociología, por lo que no resultan fáciles de leer para los profanos y los no familiarizados con el tema. Wernet (2000) ofrece una introducción más sencilla a la Hermenéutica Objetiva y al análisis secuencial. En cuanto al análisis de casos individuales, se recomienda el volumen editado por Kraimer (2000). Hildenbrand (2005, 2018) es adecuado para la aplicación concreta en la investigación familiar o el análisis de genogramas y Hildenbrand (2006a) para la conexión de la teoría fundamentada y la Hermenéutica Objetiva en el contexto de los análisis de casos individuales.

Por lo tanto, los lectores interesados encontrarán en la bibliografía aquí citada suficientes puntos de referencia para poder informarse de forma independiente. Seguiremos ahora a Wernet (2000), que se centra en la competencia metodológica de la Hermenéutica Objetiva: el análisis de secuencia. Wernet aborda la práctica concreta del análisis, los pasos necesarios y los posibles escollos de una forma fácil de entender para los principiantes que se pueden interponer en el camino hacia un conocimiento científico.

11.7 Concentración metodológica en lo esencial

En su trabajo sobre el análisis secuencial, que se centra en la práctica metodológica y a menudo se denomina análisis secuencial fino (el "fino" significa que entonces funciona aún más minuciosamente de lo que ya lo hace, por ejemplo, Studer, 1998, A3.1.2.3), Wernet (2000) ofrece un marco relativamente libre de teoría que permite analizar secuencias con un conocimiento mínimo de la teoría de la Hermenéutica Objetiva. Esto no significa que el análisis de secuencia se elimine así del contexto teórico. Sin embargo, en nuestra opinión, no es necesario comprender y aplicar plenamente la Hermenéutica Objetiva en todos sus aspectos para trabajar metodológicamente de forma limpia en un problema concreto con el análisis de secuencia y obtener el máximo beneficio de él. Incluso iríamos un paso más allá: Aunque no cabe duda de que el análisis secuencial procede de la Hermenéutica Objetiva y está estrechamente vinculado a ella, consideramos que el análisis secuencial es un formato universal para extraer y combinar información de los textos con el fin de llegar a conclusiones plausibles de una manera adecuada al caso y sensible al contexto.

En sentido estricto, lo llamaríamos la contrapartida puramente cualitativa de la estadística bayesiana. Ambas se caracterizan por la orientación concreta al contexto y el aprendizaje a partir de la experiencia, lo que acumulativamente repercute directamente en los resultados del análisis. Debido al alto grado de adaptación a cualquier contenido contextual ("teoría del objeto" en el discurso científico), la Hermenéutica Objetiva adquiere así el estatus de metateoría para poder trabajar sobre la base de la teoría del objeto. Vemos una cierta singularidad en el análisis secuencial, ya que de nuestra experiencia investigadora con él no conocemos hasta la fecha ningún caso en el que los datos evaluados mediante el análisis secuencial hayan requerido un análisis estadístico posterior en el sentido de los métodos mixtos. Sin embargo, conocemos estudios (Studer, 1998) que combinan los resultados de la Hermenéutica Objetiva con rigurosos análisis estadísticos de otros aspectos del objeto de estudio y pueden producir así una imagen muy vívida e impresionante de un fenómeno.

En cuanto a la calidad de los resultados de los análisis secuenciales, sus resultados – siempre que el trabajo metodológico sea limpio – nos resultan tan convincentes que podemos proceder directamente a la puesta en práctica y aplicación de las conclusiones. También estamos familiarizados desde la práctica de la terapia de adicción (Gürtler, Studer & Scholz, 2012) o la gestión de casos (Studer, 2005) y desde varios talleres y seminarios con el análisis de casos de forma adecuada con legos científicos en un entorno de grupo. Si se respeta cuidadosamente el control metodológico, no se observan pérdidas significativas en la calidad del análisis. Si se observan las reglas pertinentes (véase el capítulo 11.9) y las fuentes de error (véase el capítulo 11.9.6), es muy posible llevar a cabo análisis muy aceptables y orientados a la práctica con no científicos sin que cada una de las personas presentes tenga que entender o incluso ser capaz de desenvolverse en todo el trabajo de la Hermenéutica Objetiva. Sobre la base de estas experiencias, existe un alto grado de validez ocular de que el método analítico estrechamente definido del análisis de secuencia se puede aplicar ciertamente sin la Hermenéutica Objetiva como teoría de fondo, pero respetando estrictamente las reglas del análisis. Esto tiene dos consecuencias para el debate posterior:

- Concentración en la práctica metodológica del análisis secuencial.
- Discusión de la inserción del análisis secuencial en el contexto de otros métodos, especialmente MixedMethods.

Sin embargo, antes de llegar al conjunto concreto de reglas, empezamos con un estudio de caso.

11.8 Un estudio de caso de la práctica terapéutica

La práctica metodológica es más fácil si se demuestra con un estudio de caso. Un ejemplo breve y muy bien elaborado de la práctica cotidiana de la terapia de la adicción procede de Studer (1995), que analiza una carta de solicitud de un cliente potencial para la terapia de la adicción. Para otros análisis, por ejemplo de genogramas y biografías, remitimos a Hildenbrand (2005,2018) y para otros análisis de cartas de solicitud y biografías a Gürtler, Studer y Scholz (2012).

Una formulación frecuente de clientes potenciales de terapia de adicción respecto a su motivación de terapia pegadiza es a saber (Studer, 1995, p.19f.):

21 Todavía estoy muy interesado
22 en conseguir por fin
23 controlar mi drogadicción. ...

Podemos leer e interpretar este segmento de texto (es un extracto de un texto más largo) como

- "Quiero liberarme de la adicción" o
- "Quiero poder vivir de forma abstinente", pero también
- "Los intentos anteriores han fracasado, pero sigo queriendo" y, en consecuencia
- "Quiero (aprender) a ser capaz de controlar mi consumo de drogas" – porque lo que controlas, no lo sueltas.

Se puede especificar esto de la siguiente manera

- "Quiero consumir, pero por favor sin sufrir efectos secundarios".

Del "todavía" se puede hacer la hipótesis: "Este no es mi primer intento". ¿Cuántos intentos ha habido antes? ¿Estamos ya ante un cliente "fuera de terapia"? ¿Qué nos espera entonces en el proceso terapéutico y en la vida cotidiana de un entorno hospitalario con esta persona?

Como podemos ver, estas hipótesis conducen más o menos directamente a otras preguntas y conclusiones. El pasaje "gran interés" contrarresta la necesidad existencial, que presumiblemente es "Si no cambio pronto, puedo morir" o incluso "... moriré con una alta probabilidad". El miedo existencial por la propia existencia, es decir, por poder seguir viviendo en condiciones humanas lejos de la violencia, la prostitución y, en general, la presión diaria de procurarse drogas, no puede ser un "interés". Es un "deber", una apelación a la propia voluntad e instinto de supervivencia, ya que la persona puede estar ya cerca de un "o" respecto a su propio modo de vida. En este punto, los matices de gris y las gradaciones en la vida desaparecen cada vez más. En realidad, debería ser "¡Ayúdame, porque no puedo hacerlo solo!". En cambio, "el interés" es fingido, como si se tratara de visitar una exposición interesante o de ver una determinada película de cine – interesante, al fin y al cabo. Tampoco se trata de vivir un encuentro o un acontecimiento interesante.

En un análisis abreviado, la propia situación de la clienta potencial se presenta como si tuviera la opción de no tener que orientarse según las necesidades de su propia situación vital actual y, de este modo, intenta comunicar y representar al mundo exterior una pseudoautonomía que, de facto, no tiene y, desde luego, no ha tenido recientemente. Al mismo tiempo, intenta crear una imagen positiva de sí misma, pero esto fracasa porque, en el caso de esta clienta potencial, la motivación inicial "para controlar la adicción" se dirige hacia el control, pero no hacia la superación completa y el abandono de la adicción. El objetivo del control es evitar que ella sufra daños. Por lo tanto, no se trata de liberarse de las drogas o de autotransformarse en una vida más sana sin drogas, sino de consumir sin tener que aceptar los "efectos secundarios" perjudiciales. Sigue tratándose del consumo, sólo que de cambiar sus condiciones, no de la cuestión de "si en absoluto". Pero la terapia debería ser exactamente el lugar donde se promueve la autotransformación y la pseudoautonomía se transforma en auténtica autonomía real. Un primer paso sería la comprensión de la propia pérdida de control. En el proceso, la clienta potencial tropieza consigo misma. Por un lado, no se ofrece de forma convincente como cliente, por otro, no está en condiciones de establecer condiciones ni siquiera de formular expectativas o fingir interés cuando se trata de su propia supervivencia. A modo de comparación: en el Titanic (véanse los capítulos 5.5.4 o 12.11.2) no se trataba de interesarse por una plaza en un bote salvavidas, sino que ése era el nivel mínimo para poder sobrevivir. En consecuencia, la clienta potencial fracasa dos veces, lo que probablemente describe muy bien su situación vital en este momento.

Estas explicaciones podrían elaborarse con mucho más detalle a partir de esta única frase. Sin embargo, ya debería quedar claro que el análisis detallado de incluso una sola frase ya permite tantas interpretaciones y lecturas acumuladas que se pueden formular algunas hipótesis plausibles para el curso posterior (por ejemplo, cómo continúa la carta de solicitud) (ejemplos en Gürtler, Studer & Scholz, 2012). Además, se podrían formular muchas hipótesis sobre qué problemas trae realmente la clienta a la terapia y cómo se posiciona la institución para ayudar a la clienta en la medida de lo posible a desprenderse cada vez más de la drogadicción y a llevar su propia vida en una dirección más curativa.

En el caso que nos ocupa, cabe suponer que la clienta potencial no está realmente interesado en la terapia y, por lo tanto, sólo se la tomará en serio hasta cierto punto. Esto significa que tendrá problemas con ciertas normas y no las cumplirá o intentará socavarlas. Es probable que consuma, sólo para probar si funciona (ahora) sin "efectos secundarios perjudiciales". Asimismo, es posible que ya esté "bien o por lo menos al fin de las posibilidades terapeúticas", ya que no es su primer intento. Esto significa que puede tener tendencias especialmente inconscientes a la subversión en cualquiera de sus formas, que se manifiestan especialmente en entornos terapéuticos. También puede significar que sabe intuitivamente "cómo funcionan los terapeutas" y "a qué responden". Por lo tanto, siempre es necesario examinar detenidamente qué significa para ella la autenticidad, para que pueda sentirla, percibirla y aprender a distinguirla por sí misma en primer lugar. Esta reflexión será un gran reto para ella y requiere un alto grado de persistencia por parte del terapeuta. Estos puntos ya pueden abordarse abiertamente durante una entrevista para transmitir claridad desde un punto de vista institucional y casi como efecto de primacía y sentar las primeras modestas bases para una necesaria postsocialización (Studer, 1998).

En la práctica, esto significa, como en el ejemplo que nos ocupa en la fase previa a una entrevista de admisión para una terapia de adicción en régimen de hospitalización, que los clientes pueden ser atendidos de forma adecuada a su caso. Un análisis previo muestra qué recursos y déficits tiene una persona, qué posibles obstáculos y retos puede plantear una terapia a largo plazo, etc. A continuación, se puede elaborar un plan de terapia para el cliente. De este modo, un plan de terapia puede adaptarse de forma centrada en

el cliente sin anular un concepto organizativo general (Studer, 1995, 1998). Es importante que la nueva información – por ejemplo, cuando las personas cambian, muestran nuevas facetas de sí mismas, se abren, prueban algo nuevo, etc. – se integre en la hipótesis de la estructura del caso, aunque al principio pueda parecer contradictoria. Por cierto, una discusión totalmente abierta sobre los recursos y los déficits no carece de tacto ni es peyorativa, sea cual sea el tema. Los esfuerzos se dirigen a ayudar a los clientes y para ello es necesario que no se omita ningún tema y que se aproveche al máximo el espacio de posibilidades. *Los tabúes están fuera de lugar aquí y sólo entorpecen*. No se trata de ser amable, sino de ayudar realmente a los clientes. El análisis y la comunicación van por separado. La forma de dar feedback a los clientes debe ser la adecuada para el cliente. Por esta razón, los afectados no deberían estar presentes durante un análisis, ya que muy pocos podrían soportar esta apertura durante el análisis, lo que a su vez perturbaría el proceso de análisis y en la mayoría de los casos no sería útil para nadie. Analizar sin las personas afectadas no es en absoluto una falta de respeto hacia ellas, ya que después se puede planificar cuidadosamente cómo y de qué forma o ritmo se comunican los resultados del análisis a los clientes y así se puede enmarcar y guiar el proceso terapéutico. No se desecha nada del proceso de análisis.

La tendencia humana, y desde luego la de científicos y profesionales, es a menudo reinterpretar la información nueva o incongruente, bloquearla e ignorarla, o presentarla como irrelevante. El análisis secuencial en el marco de la Hermenéutica Objetiva permite evitar todos estos errores o reducirlos al mínimo si se aplica correctamente. Así, la información se pondera por igual, nada se olvida, nada se exagera; y sólo una integración de toda la información y no la preferencia de un subconjunto seleccionado arbitrariamente conduce a una imagen global plausible de la información disponible. De todos modos, desde un punto de vista científico, es aconsejable contemplar el mundo en términos de *información*: ¿qué significa la información y de qué manera marca la diferencia (Bateson,1985)?

El conocimiento de las estructuras de significado latentes reconstruidas de este modo aporta la inestimable ventaja para la práctica de estar cerca de la realidad de los clientes, que repercute en sus acciones y determina su vida cotidiana. En primer lugar, hay que tomar en serio y valorar su motivación expresada en el momento de cambiar su modo de vida anterior. Sin embargo, a esto le sigue una exploración realista del marco de posibilidades para poder abandonar a tiempo las ilusiones terapéuticas y adaptar individualmente las ofertas terapéuticas.

Para otros contextos (educación, empresa, supervisión y coaching, selección de personal, formación, desarrollo de talentos, etc.) se aplican directrices comparables y adaptadas al contexto.

Para ilustrar esto, nos gustaría citar otro segmento de texto de otra carta escrita por un cliente (Studer, 1998, p.26), que surgió en el transcurso del trabajo terapéutico y las reflexiones biográficas:

"Pero también nosotros [mi pareja y yo], sin embargo, tuvimos muy buenas conversaciones...".

Podríamos interpretar este segmento como otro ejemplo ("Pero *también* teníamos...") del aprecio de la escritora por su pareja. El "nosotros" común podría ser un énfasis de esto. Sin embargo, podemos leer igualmente al contrario ("*Pero* también teníamos ...") que la escritora tiene dos pensamientos diferentes en su cabeza al mismo tiempo: "Por un lado mi pareja me molestaba, pero por otro lado también podíamos tener una buena conversación." Sin embargo, como muestra el análisis posterior de la estructura latente, la persona que escribe no mantiene claramente separados estos dos pensamientos y mezcla estas intenciones de acción separadas de acercamiento frente a distanciamiento. En el trabajo práctico con la clienta, la separación de los dos pensamientos abre a su vez el espacio de posibilidades para descubrir cosas nuevas sobre ella misma e integrarlas dialécticamente, es decir, para alcanzar un alto nivel de adaptación. Posteriormente, esto puede desembocar en acciones reales y se hace posible probar cosas nuevas.

Pero primero hay que entender qué es lo que realmente está en juego en ese momento antes de comenzar las intervenciones terapéuticas o de otro tipo. ¿Qué realidad accionable hay detrás de las expresiones verbales observadas? ¿Qué motiva la acción de un momento a otro? Estas preguntas deben abordarse con éxito antes de sacar conclusiones precipitadas. Otra cosa es lo que los clientes o los "afectados" por el análisis hagan con esta interpretación. Dado que los análisis no conocen tabúes, esto debe comunicarse adecuadamente en todas las circunstancias para poder hacer justicia a las distintas personas y no ofenderlas innecesariamente.

Así pues, no sólo hay que tener en cuenta la experiencia única de las personas, sino también sus realidades objetivas, que pueden aprovecharse como pautas de actuación individuales, sociales y culturalmente arraigadas. La estricta orientación hacia lo que hay en la realidad guía el enfoque de la Hermenéutica Objetiva. El punto de partida es el caso general, que se hace público y generalmente accesible mediante la reconstrucción de la estructura latente del significado. A continuación, la manifestación concreta, la forma de expresión puede ser determinada por la persona empírica y su subjetividad.

Tras este breve ejemplo de caso, pasamos a los principios prácticos del análisis de casos. Si bien, por un lado, el análisis secuencial puede aplicarse a cualquier contexto, por otro, se caracteriza internamente por un procedimiento estrictamente controlado desde el punto de vista metodológico. A continuación examinaremos este aspecto con más detalle.

11.9 Práctica metodológica del análisis de secuencia

El análisis de secuencias puede dividirse en dos pasos fundamentales: la generación de hipótesis y la comprobación de hipótesis:

1. Generación de hipótesis: la fase de construcción creativa (véase el capítulo 11.9.1).
2. Comprobación de hipótesis: la fase de comprobación crítica del texto (véase el capítulo 11.9.5).

... y exactamente en este orden. Todo el trabajo tiene lugar sobre el texto y no está desligado de él y no se desliga de él – independientemente de lo que constituya exactamente un texto en el caso concreto.

En el curso de la generación de hipótesis, se genera secuencialmente una hipótesis preliminar de estructura de caso *hasta que la estructura se repite una vez* y resulta improbable obtener más información. Llegados a este punto, no debería quedar ninguna hipótesis en competencia, sino sólo una, cuya viabilidad se comprueba posteriormente.

En el sentido del examen crítico sensu Popper, esta prueba de viabilidad se realiza de nuevo sobre el texto. Ahora, sin embargo, entran en juego partes del texto que aún no han sido objeto de análisis. La hipótesis preliminar de la estructura del caso se somete a una prueba de falsabilidad, lo que significa que debe formularse de tal manera que pueda fallar en principio. De este modo se garantiza que, en caso de probarse, surja realmente el conocimiento, ya que sólo la capacidad potencial de fracasar contribuye a una ganancia sustancial de conocimiento, ya que de lo contrario prevalecería una estrategia de inmunización.

Si la hipótesis de la estructura de casos propuesta supera con éxito estas pruebas en lugar de fracasar debería verse reforzada por las instancias críticas. Los puntos críticos son explícitamente como un intento dirigido de falsación. No en vano se denominan: ¡comprobación crítica! Si la teoría se demuestra a sí misma, puede ser tratada directamente como el resultado final del análisis.

Por supuesto, una nueva información – posterior al análisis – podría conducir a una revisión del resultado y a la falsación. En ese caso, habría que revisar o ampliar al menos partes del análisis y volver a examinar el material de datos a la luz de esta nueva información. La calidad de un análisis depende de la información disponible.

Si se rechaza la hipótesis de la estructura de casos propuesta (es decir, se falsifica), habrá que volver a examinar todo el proceso de análisis, incluidas todas las subhipótesis, etc., para localizar el error aparentemente presente. Dicho error (véase también el capítulo 11.9.6) estará presente a menudo. Puede ser que no se hayan respetado las características del análisis (véase el capítulo 11.9.2) – por ejemplo, la literalidad -- o que las preferencias subjetivas hayan guiado el análisis – por ejemplo, que se haya pasado por alto o clasificado como no importante determinada información del texto en lugar de argumentos verificables intersubjetivamente. Si se detecta el error, a partir de ese momento hay que "olvidar" todos los pasos posteriores existentes y continuar como si fueran "nuevos". El procedimiento no cambia. Se trabaja hasta la repetición completa de la estructura y luego se vuelve a la crítica de la hipótesis preliminar de la estructura del caso. Básicamente, este proceso puede repetirse tantas veces como se desee y es muy similar al teorema

de Bayes (véase el capítulo 6.4), que permite calcular una y otra vez un estado de conocimiento actualizado ante la aparición de nueva información. Esto se aplica independientemente de si esto cambia mucho, poco o nada las conclusiones. Sin embargo, si la hipótesis preliminar de la estructura del caso se falsifica varias veces seguidas, el problema es más profundo y los investigadores deben buscar supervisión experimentada, ya que es posible que el procedimiento en sí no se haya comprendido y aplicado completamente.

A continuación se presentan los detalles de cada paso y sus características. Lo bueno, desde nuestro punto de vista, es que hasta ahora *siempre* hemos experimentado que cuando se trabaja de forma limpia y metódicamente controlada, del análisis surge algo plausible y razonable. Esto promete una ganancia teórica de conocimiento o conduce a derivaciones claras y recomendaciones para la acción en la práctica, independientemente del contexto (terapia, educación, empresa y recursos humanos, ... deporte, ...). Incluso utilizamos el análisis de secuencias en un contexto privado un poco "relajado" (es decir, no llevado a cabo de una forma tan estricta y lenta) para encontrar a los médicos adecuados a través del análisis de páginas web o para comprobar nuestros propios correos electrónicos para ver lo que realmente estamos diciendo aquí – antes de enviarlos. Por supuesto, funciona igual a la inversa: ¿qué intenta decirme alguien en una carta, un correo electrónico, etc.? Muchas otras aplicaciones son concebibles y útiles.

11.9.1 Generación de hipótesis

11.9.1.1 Sobre el fondo teórico

> "El objeto concreto de los procedimientos de la hermenéutica objetiva son los protocolos de acciones o interacciones sociales reales, simbólicamente mediadas, ya sean escritas, acústicas, visuales, combinadas en diversos medios o fijaciones archivadas de otro modo" (Oevermann, Allert, Konau & Krambeck, 1979a, p.378).

En estos protocolos hay que encontrar el significado latente: éste es el objetivo de la Hermenéutica Objetiva. Según Wernet (2000, p.39.), el análisis sigue una secuencia consistente en

1. "Contar historias" – historias en las que podría ocurrir el segmento de texto,
2. "formar lecturas", es decir, ordenar los relatos comparativamente según semejanzas y diferencias, y
3. "confrontar estas lecturas con el contexto real", es decir, comprobar sistemáticamente la validez del texto.

No sólo hay que tener en cuenta la experiencia única de las personas, sino también sus realidades objetivas, que pueden aprovecharse como pautas de acción individuales, sociales y culturalmente arraigadas. La estricta orientación hacia "lo que es", hacia la realidad, guía el procedimiento en la Hermenéutica Objetiva. El punto de partida es siempre *el caso general y normal*, que se hace público y accesible mediante la reconstrucción de la estructura latente del significado. Después, la manifestación concreta, la forma de expresión puede ser interpretada con mayor precisión por la persona empírica y su subjetividad. Oevermann (2002, p.33) afirma:

> "El análisis secuencial anida en su estructura básica en los acontecimientos humano-sociales reales y, por lo tanto, no es, como los métodos de medición y clasificación, por lo demás habituales, un método externo al objeto, sino uno correspondiente y apropiado para la cosa misma. De hecho, en la vida práctica, las decisiones también deben tomarse en principio en cada punto de la secuencia entre las opciones aún abiertas a un futuro abierto."

El análisis secuencial "[...] se apoya en la secuencialidad constitutiva de la acción humana" (ibíd., p.6). Sin embargo, un enfoque secuencial no significa simplemente trabajar de adelante hacia atrás. Se trata más bien de abrir nuevas posibilidades de interpretación ("lecturas") a lo largo de la estructura natural del texto y volver a cerrarlas cuando no pueden resistir el escrutinio del texto.

Repetimos: el análisis de secuencias es la interacción entre *la posibilidad* (lo que podría ser) y *la realidad* (lo que es realmente compatible con el texto).

11.9.1.2 Principios de la generación de hipótesis

Como ya se ha citado, Wernet (2000, p.39) da una sencilla "respuesta a la pregunta: ¿qué tengo que hacer para llevar a cabo una operación metodológicamente verificable de reconstrucción del significado según reglas válidas?". La respuesta consiste en un proceso metodológico de tres pasos, de los cuales los dos primeros nos interesan aquí para la generación de hipótesis: (1) contar historias, (2) formar lecturas. El tercer paso – (3) confrontar las lecturas con el contexto real – tiene lugar en un doble sentido. En primer lugar, el conjunto de lecturas e interpretaciones potencialmente válidas se confronta repetidamente con cada nueva parte del texto, lo que conduce a la formación de la posterior hipótesis preliminar de la estructura del caso. En segundo lugar, en la fase de examen crítico (véase cap. 11.9.5), esta hipótesis provisional de la estructura de caso se somete específicamente a un examen crítico, es decir la hipótesis de la estructura de casos se somete deliberadamente a falsación. Por naturaleza, a efectos analíticos, primero nos concentramos en la generación de hipótesis antes de poder comprobarlas.

Antes de empezar a interpretar segmentos de texto, es necesario aclarar *qué es el caso* y *en qué contexto está incrustado*. Según Wernet (2000), esto incluye, por un lado, aclarar y revelar el interés de la investigación y, por otro, aclarar qué es lo que el protocolo de texto registra realmente, qué realidad social registra o, más concretamente, qué contribución puede hacer una entrevista, por ejemplo, para responder a las preguntas de la investigación.

El primer paso empieza con los datos objetivos del caso (nacimiento, entorno, ocupaciones, fechas de fallecimiento, etc.), y sólo después se somete a análisis el texto propiamente dicho (conversaciones en clase, conversaciones terapéuticas, entrevistas biográficas, etc.). La recogida de datos objetivos corresponde, por ejemplo, a la elaboración de un genograma (datos familiares). Sin embargo, dependiendo del contexto, estos datos sólo están disponibles de forma limitada, por ejemplo, en contextos educativos, empresariales/ personales, etc.

11.9.2 Reglas de interpretación

La interpretación propiamente dicha se rige por las cinco reglas de *libertad del contexto, literalidad, secuencialidad, extensividad o totalidad y parsimonia* (Wernet, 2000, pp.21-38):

11.9.2.1 Libertad del contexto

Por supuesto, la interpretación de un texto debe tener en cuenta el contexto en el que se escribió el protocolo, pero sólo *después* de haber desarrollado los posibles significados de un segmento del texto independientemente de este contexto. De lo contrario, el análisis se desliza con demasiada rapidez hacia profecías autocumplidas, porque entonces las interpretaciones se crean en función del contexto y no en el texto. Así pues, primero se inventan historias en las que el pasaje del texto crítico podría tener sentido. Wernet habla aquí de diseñar exclusivamente "contextos experimentales de pensamiento" para generar posibles significados del segmento de texto en el primer acceso al texto. Wernet advierte de que la interpretación dependería de la comprensión cotidiana del intérprete o de la comprensión previa no obtenida científicamente y, por tanto, podría ponerse en marcha un proceso circular de interpretación si la comprensión previa no se deja de lado metódicamente por el momento. Esto significa que, como analizadores, tenemos que entrenarnos para perder selectivamente la memoria. Del mismo modo, conviene aprender a leer sólo hasta un determinado punto del texto y no una letra más allá.

11.9.2.2 Literalidad

La literalidad significa que la interpretación se guía precisa y exclusivamente por lo que ocurre real y verificablemente en el texto, y no por lo que el texto podría haber querido decir. El texto como tal es un fragmento de la *realidad* y debe tratarse como tal. El texto constituye una referencia absoluta. De ello se derivan dos condiciones importantes:

1. El protocolo del texto debe – en realidad, como algo natural – constituir la base de la interpretación en su forma original, por ejemplo, como transcripción palabra por palabra de una interacción social, y no en forma de paráfrasis alienada por el entendimiento cotidiano y las convenciones sociales y ya reducida de antemano en sus posibilidades de significado. Tales paráfrasis, tal como se utilizan a veces en los análisis cualitativos de contenido, están absolutamente fuera de lugar. Las paráfrasis son una reliquia de la era pre-informática y no tienen ningún significado real en la actualidad. Si no se trata de paráfrasis como tales, no desempeñan ningún papel en el análisis de datos, ya que distorsionan los datos originales y los interpretan al mismo tiempo.
2. La interpretación debe ser francamente exigente con el texto. No debe pasar por alto detalles que parezcan accidentados o incluso inapropiados, como podríamos hacer en la vida cotidiana. Wernet (ibíd.) exige que el texto sea "sopesado en la balanza" de un modo que "parecería mezquino" en la vida cotidiana. Encontramos esta capacidad de "concretismo" especialmente pronunciada en los niños pequeños, cuya percepción a menudo aún no ha sido tan fuertemente distorsionada por su socialización y los tabúes, barreras e interpretaciones habituales que ésta conlleva. Podemos aprender mucho de ellos.

De este modo, inapropiado en las conversaciones cotidianas y quizá incluso escandalosamente beckmesseriano, se revela el significado latente tras la formulación manifiesta. En los ejemplos anteriores "Quiero controlar mi adicción" y "Pero también tuvimos muy buenas conversaciones" la interpretación literal encuentra la desviación de lo que se quiere decir con respecto a lo que realmente se dice y, por tanto, la ambigüedad de estos enunciados y las discrepancias en el mundo vital de la persona.

11.9.2.3 Secuencialidad

Este principio determina el núcleo del análisis, ya que el procedimiento de interpretación es estrictamente lógico y está orientado paso a paso. Esto hace que el procedimiento sea científico. Así, no se buscan pruebas de posibles interpretaciones (lecturas) de forma transversal y no sistemática a través del texto (búsqueda de confirmación positiva). Más bien, se alterna la formulación de hipótesis, su condensación en lecturas y su cotejo con el texto en una secuencia estricta. Esto demuestra que los análisis cualitativos se enfrentan en principio a los mismos problemas que los cuantitativos cuando se trata de conclusiones erróneas (véase la Tab. 4.2, p. 78).

Dónde comienza exactamente la interpretación y qué segmento del texto representa la secuencia inicial debe justificarse en términos de contenido. No es obligatorio empezar por la primera frase o parte de un segmento de frase. Sin embargo, si se opta por un segmento de texto, a partir de ese momento el trabajo sigue metódicamente estricto, paso a paso a lo largo de la secuencia natural del protocolo. Lo que sigue en el texto tras el segmento actualmente en el punto de mira no debe tenerse en cuenta en un primer momento. Por supuesto, cada segmento de texto está incrustado en el "contexto interno" del significado del texto que se ha interpretado. El hecho de no tener en cuenta los segmentos de texto que siguen es muy importante desde el punto de vista metodológico y práctico (Wernet, 2000, p.30): "La progresión pensamiento-experimento deja claro que el respectivo caso concreto debe tomar la 'decisión' de ser lo que es...". Wernet subraya en el mismo pasaje que esto "...se refiere a la reconstrucción del 'ser-como-es' de una práctica vital".

El principio de secuencialidad no prohíbe saltarse segmentos de texto y seleccionar pasajes relevantes del texto en función de la pregunta de investigación y del progreso de la reconstrucción de la estructura de sentido del texto. La selección de los pasajes de texto y el trabajo secuencial dentro de ellos deben estar separados entre sí. Sin embargo, el inicio del análisis debe justificarse de nuevo y la interpretación debe proceder de nuevo secuencialmente paso a paso. La selección de pasajes de texto posteriores se basa en si pueden apoyar o incluso confrontar las distintas lecturas.

Secuencialidad significa también no dejar que fluyan en la interpretación conocimientos previos sobre desarrollos posteriores del texto, sino orientarse al desarrollo paso a paso en la interpretación. Los analistas

asumen así una ignorancia e ingenuidad elegidas intencionadamente. En concreto, esto significa ser consciente de las cosas que pertenecen al caso. Pero este conocimiento no se utiliza para apoyar una determinada lectura ni para debilitar otra en el sentido de una confirmación positiva. Por eso es mejor trabajar en grupo, ya que así se reduce la propia ceguera operativa y todos se comprometen a trabajar según las reglas analítico-secuenciales. Aparte de los desarrollos potencialmente eficaces de la dinámica de grupo, que pueden ocurrir en grupos pequeños (Stoner, 1961), esto en realidad siempre resulta útil y necesario en la práctica.

Así pues, el trabajo secuencial permite la adquisición de dos competencias particulares:

1. conocer la información en principio, pero no utilizarla explícitamente, para no anticiparse y cometer errores metodológico-lógicos.
2. leer las frases de un texto sólo hasta donde llega la unidad de análisis y no más allá, para no cometer el mismo error de utilizar la información de forma inadecuada.

De este modo, no se debe utilizar para el análisis información que todavía se desconoce en este punto debido al procedimiento secuencial. En la práctica, ayuda simplemente tapar el resto del texto. Entonces ya nadie podrá leerlo. Sólo se descubre tanto como se analiza. El uso del ordenador (s. cap. 11.12 o A), por ejemplo, con AQUAD 7 (Huber, 2019) el único programa informático disponible en la actualidad que ha implementado el análisis de secuencia, sólo muestra el fragmento de texto a analizar. Esto significa que este problema no se produce en absoluto. Desgraciadamente, uno no aprende esta maravillosa habilidad del enmascaramiento selectivo.

Repetimos: no sólo es importante buscar la confirmación de las hipótesis, sino más bien buscar pruebas en contra para poder refutarlas. Esto se corresponde con el enfoque falsificador del racionalismo crítico (Popper, 1943). La reconstrucción del significado debe ceñirse exactamente al texto y verificarse exactamente allí. En la práctica, sin embargo, es posible identificar intentos inconscientes de apoyar empíricamente las propias ideas modelo en lugar de cuestionarlas. Entonces, el examen imparcial de las hipótesis en liza sólo tiene lugar de forma limitada. Esto es similar al problema esbozado por Gelman y Loken (2013) de que el diseño de los estudios a menudo se orienta hacia la consecución de los resultados esperados o que se disponga a posteriori de muchas explicaciones para los datos encontrados (cf. datos a posteriori; véase en el capítulo 4.4.2.2, una crítica del estudio de Bem). Tales errores no sólo ocurren necesariamente en el análisis, sino ya en el diseño. La documentación transparente de las actividades científicas desempeña un papel que no debe subestimarse para la detección de tales tendencias.

11.9.2.4 Extensividad o totalidad

La extensividad o totalidad es un principio de equilibrio que debe impedir que se sigan intenciones subjetivas y arbitrarias a la hora de interpretar. Así, no deben pasarse por alto detalles aparentemente sin importancia y la interpretación no debe limitarse a los pasajes aparentemente importantes. No interpretamos en la línea de "Oh, no quisó decir eso, podemos ignorarlo", sino en la línea de "Veamos qué viene ahora... y ahora... y ahora...".Importante y no importante son consecuencias directas de prejuicios subjetivos. Por lo tanto, se procede en la interpretación de tal manera que el texto se analiza paso a paso sin favorecer ni desfavorecer nada. Es esencial tratar todos los pasajes por igual. El objetivo es explorar a fondo las posibles interpretaciones y no reducirlas.

11.9.2.5 Parsimonia

Este principio se basa en la navaja de Occam, un criterio de la filosofía de la ciencia que se remonta al filósofo y teólogo Guillermo de Ockham (1288-1347) en la escolástica tardía. Se refería principalmente a la prueba aristotélica:

Recordatorio 11.2: La navaja de Ockham

Por tanto, es preferible la teoría que pueda expresar lo mismo con el mismo alcance en términos más sencillos y que requiera menos presuposiciones o contenga menos variables. La claridad y la comprensibilidad lógica también se aplican a las referencias de los elementos de la teoría entre sí.

El principio exige que se asuma la normalidad, entendida como lo cotidiano con sus rutinas, a la hora de construir posibles interpretaciones de la práctica vital. La exigencia aquí, según Wernet (ibíd.), es permitir sólo interpretaciones (lecturas) que puedan verificarse en el texto. Esto limita la narración a variantes compatibles con el texto en su conjunto y excluye por el principio de parsimonia los relatos de ciencia ficción, las divagaciones esotéricas o, sobre todo, las atribuciones infundadas pero siempre excitantes de desviaciones patológicas del comportamiento normal.

La desviación de lo normal debe revelarse a través del texto y no se espera como norma. Una desviación de la norma, que a menudo se considera patológica, destaca muy fácilmente en los protocolos. Esto se debe a que aquí el margen de interpretación se limita drásticamente. Las rutinas y las soluciones sostenibles a los problemas llaman menos la atención porque el espectro de posibilidades es muy amplio. Por tanto, la normalidad es más difícil de reconstruir que la desviación de la norma.

La frugalidad debe entenderse de dos maneras:

- Por un lado, la normalidad se da inicialmente por supuesta y la asunción de la desviación está sujeta a la obligación de justificar. Se trata de descubrir desviaciones realmente significativas de la normalidad de la acción humana.
- Por otro lado, el espacio de posibilidades se ve limitado por el hecho de que la interpretación sólo se realiza sobre el texto existente y todas las conclusiones se basan únicamente en el texto, pero no en experiencias cotidianas posteriores.

En resumen, la máxima es: lo que se interpreta debe fundamentarse en el texto y lo que se fundamenta pertenece a la interpretación.

11.9.3 Condensación de hipótesis en lecturas del texto

La totalidad de las reglas del análisis de secuencia, que generan una multiplicidad de opciones de nuevo en cada punto de la secuencia y en cada segmento del texto, forma un conjunto de reglas tipológicamente diferentes. Como ejemplos de estas reglas, Oevermann (1996a, p.7) menciona la sintaxis lingüística, las reglas pragmáticas de la acción del habla, así como las reglas lógicas para el razonamiento formal y material-sustantivo (es decir, inducción, abducción, deducción, véase el capítulo 2; Reichertz, 2000). Oevermann (1996a) habla de enunciados bien formados con respecto a las interpretaciones que surgen de las posibilidades.

Existe un enunciado bien formado cuando se ha generado una interpretación de acuerdo con las reglas y con los bloques de construcción de una lengua, cuando es un enunciado lingüístico normal escrito u oral que se ajusta a todas las reglas de la lengua en cuestión en cuanto a elección de palabras, estilo y gramática. Se trata, pues, de un enunciado gramaticalmente correcto y, por lo tanto, de una frase completamente normal. En realidad, esto podría expresarse más fácilmente y sin pérdida de precisión en lugar de envolverlo en un complicado alemán sociológico.

Los relatos generados hasta ahora se agrupan para derivar tipos. En su formulación, estos tipos forman las lecturas del texto. Sólo entonces se contrastan las lecturas con el texto y se confrontan con la realidad del

texto. De este modo, la conexión entre lo general (estructura abstracta e independiente del caso) y lo particular (práctica vital concreta) se esboza con mayor claridad y precisión. Técnicamente hablando, como ya se ha explicado varias veces, se aplica el principio de falsación. Esto significa que un supuesto no se demuestra, sino que se mantiene, porque su contrario, su inaplicabilidad al caso concreto, no puede demostrarse. Esta es la tradición de Popper. En consecuencia, no existe una verdad siempre válida, sino sólo hipótesis demostradas cuya no-demostrabilidad aún no ha sido demostrada. Sin embargo, recordemos: puesto que siempre podemos equivocarnos, también podemos equivocarnos al falsar. Una falsación pertenece al escrutinio crítico tanto como una confirmación. Por decirlo de un modo más sencillo: como sólo tenemos información de nuestro universo (como quiera que lo definamos), sólo podemos hacer afirmaciones limitadas sobre él. La información procedente del exterior de nuestro universo sería completamente significativa. Pero esto no está disponible y, por tanto, cualquier discusión sobre la extensión de la verdad relativa tras una verdad absoluta en la ciencia termina bastante rápido. El resto es una cuestión de fe y puede seguir investigándose en las religiones, pero entonces ya no es científico.

11.9.4 Generación de una hipótesis proposicional de estructura de casos

Puesto que el objetivo general del análisis es la generación de una estructura de casos, nos remitiremos a otra definición de este término según Hildenbrand (1996). Hildenbrand (ibid., p.6) define la estructura de casos como

> "la manera casuística-regular-habitual [de una persona o instancia de acción] de ver, interpretar e intervenir en el mundo, en pocas palabras: de construirlo como significativo ...
> Este contexto de significado que forma un caso y que lo caracteriza en su estructuración tiene múltiples capas. Alcanza su máxima complejidad en el lenguaje, que representa por tanto la fuente central de material en la reconstrucción de casos. Pero también se materializa en otras cualidades expresivas que pueden describirse y, por tanto, textualizarse. Ejemplos de ello son el comportamiento no verbal, el mobiliario doméstico, el simbolismo privado, etc.".

La generación de la hipótesis de estructura de caso propuesta, que luego se somete a prueba crítica en el texto de acuerdo con Popper se basa en un desarrollo evolutivo:

- Al principio surgen muchas posibilidades de interpretación y lecturas posteriores.
- Estas se estrechan drásticamente a través de la creciente ganancia de información y a través de la creciente comprensión del texto y su secuencia natural, es decir, de acuerdo con un procedimiento estrictamente secuencial.
- Este estrechamiento se practica hasta que sólo queda una lectura global con una alta complejidad de contenido.
- Esta lectura final corresponde a la hipótesis preliminar de la estructura del caso. Sigue teniendo carácter de hipótesis, ya que falta el examen crítico del texto. Sin embargo, siempre que el trabajo metodológico sea limpio, ha pasado por un proceso evolutivo exitoso y ha demostrado ser superior a todas las demás hipótesis o lecturas competidoras, además de haber integrado sus contenidos.

Si se infringen las normas de interpretación del texto, pueden seguir diversas consecuencias. Por ejemplo, puede que una hipótesis audaz no se exprese y se lleve a cabo realmente, aunque sea precisamente la posibilidad potencial del fracaso de una suposición lo que tiene el poder explicativo. Una hipótesis es fuerte si asume el riesgo de ser derribada incluso por una pequeña discrepancia. De ello se deriva un mayor ámbito de validez. Una hipótesis que no pueda refutarse en principio carece de valor científico.

Hay que investigar todas las posibilidades, por pequeñas que sean, de entender un texto. Naturalmente, este requisito hace que el método sea muy costoso, pero conduce al hecho de que una muestra muy pequeña es suficiente para abarcar exhaustivamente un campo de investigación. En el cálculo del coste total, el análisis de secuencia sale bastante bien parado, aunque el análisis de textos individuales parezca inicialmente demasiado costoso. En el cuadro general, las cosas parecen diferentes. El procedimiento es muy eficaz. Se

analiza *hasta que la estructura del caso se reproduce completamente una vez* (Oevermann, 1996a). Desde la lógica interna se muestra de nuevo y repetidamente en cada sección, basta con examinar unos algunos pasajes con gran detalle para obtener, no obstante, una visión de conjunto del caso. Esta estabilidad de la interpretación textual legitima analizar sólo hasta la primera repetición completa de la estructura y, a continuación, buscar activamente contrapruebas que invaliden la estructura. Esto se pone de relieve por el hecho de que "... una estructura de caso ... sólo se conoce [realmente] si se ha reconstruido una fase completa en su reproducción o transformación" (ibíd., p.9).. El análisis da lugar entonces a las estructuras de sentido manifiestas y obvias o latentes de la rutina cotidiana en situaciones estándar (ibíd., p.76s.).

11.9.5 Comprobación crítica de hipótesis

Una vez reconstruida la estructura básica de cada segmento textual y delimitado el espacio de posibilidades, las hipótesis (interpretaciones, lecturas) generadas en el proceso se someten a prueba de falsedad en el siguiente segmento textual. Probar una hipótesis equivale, por tanto, a intentar demostrar y justificar que la hipótesis no es compatible con el texto. Pueden darse los siguientes casos

- Una hipótesis es congruente con la parte textual sin necesidad de ninguna otra modificación. Se mantiene.
- Una hipótesis no es congruente con la parte textual, pero una modificación la hace congruente. Se mantiene en la forma modificada.
- Una hipótesis no es congruente con la parte del texto y ninguna modificación adicional permite que sea congruente con la parte del texto. Se rechaza (falsa) y no se mantiene.
- Una hipótesis no existía hasta ese momento, pero surge como resultado del trabajo sobre el texto. Se incluye en el conjunto de lecturas y, posteriormente, también se la examina críticamente.

Hasta ahora, esto muestra el procedimiento en el trabajo paso a paso para crear la hipótesis preliminar de la estructura del caso. Teóricamente, podemos equivocarnos sobre cualquier caso. El trabajo en grupo nos ayuda a minimizar estos errores. Con la hipótesis de la estructura del caso generada de este modo, que por razones metodológicas debe formularse claramente por escrito en este punto, se puede *pasar al paso de la comprobación crítica de la hipótesis*. Ya no es necesario hacerlo de forma secuencial. En la búsqueda de pruebas para las historias generadas en el curso de la generación de hipótesis, es necesario *deambular* por el texto – específicamente para encontrar *precisamente* esas contrapruebas. De este modo, se utilizan específicamente pasajes del texto para falsar la hipótesis preliminar de la estructura del caso.

11.9.6 Errores típicos en la realización del análisis de secuencias

Los errores en la realización del análisis de secuencias suelen ser el resultado de violaciones de las reglas de interpretación presentadas. En principio, se deben evaluar las hipótesis en función de su potencial de fracaso. Si una hipótesis no puede fallar en principio, entonces nada sustancial puede salir de ella. En ese caso, no se puede explicar nada y cualquier esfuerzo es inútil. Si una hipótesis que en realidad podría derribarse fácilmente no se derriba mediante un examen crítico y, en cambio, resulta ser viable, tiene una alta verosimilitud.

Dividimos las posibles fuentes de error en la interpretación según las cinco reglas ya comentadas: libertad de contexto, literalidad, secuencialidad, extensividad o totalidad y parsimonia. Sin embargo, nos gustaría empezar con otro tema: la *presión interna* de querer terminar un análisis. Los psicólogos hablan del *delay of gratification (retraso de la gratificación)*. Operacionalizado, el concepto muestra la diferencia entre resistirse a una recompensa inmediata en previsión de una recompensa mayor, posterior. En experimentos con niños, se les da un chocolate con la tarea de no comérselo en los próximos 15 minutos, por ejemplo, porque entonces recibirán uno mayor o mejor. Si se comen el chocolate antes de que transcurra ese tiempo, no reciben el otro chocolate que en realidad prefieren. Y eso es exactamente de lo que se trata también.

11.9.6.1 Presión interna para terminar

Se necesita perseverancia, disciplina y constancia para poner en práctica todas las reglas. Se suele decir que se necesitan al menos entre 10 y 15 horas para interpretar entre 2 y 4 minutos de una transcripción de interacción. Este procedimiento, bastante extenuante, se nutre de elaborar una estructura de caso completa a partir de los pocos pasajes interpretados.

No cabe duda de que el procedimiento requiere mucho tiempo. Sin embargo, si se tiene en cuenta cuánto tiempo consume en comparación una investigación estadística, o cuánto tiempo requiere la codificación dentro del paradigma de codificación, el análisis de secuencia adquiere facetas mucho más eficientes. Con una cuidadosa selección de los casos (principio de máximo contraste, Hildenbrand, 1999, cap. III), de ocho a diez casos son suficientes, según Hildenbrand (2006a), para alcanzar la saturación teórica de acuerdo con la estrategia de investigación de la teoría fundamentada según Glaser et al. (1998), de modo que más casos no prometen ninguna ganancia sustancial de conocimiento más allá de eso. El propio Oevermann llega incluso a afirmar que con un solo caso bien analizado se pueden hacer afirmaciones de gran alcance. Nosotros no recomendaríamos esto. Sin embargo, la valoración de Hildenbrand parece muy plausible. Esto da en resumen, entre 10*8 = 80 y 15*10 = 150 horas de tiempo de análisis para trazar cuidadosamente todo un campo de investigación – más el tiempo de selección de casos, recogida de datos, transcripción, etc. 150 horas son unas buenas cuatro semanas de trabajo. Desde un punto de vista académico, parece un tiempo razonable para un trabajo de mayor envergadura.

En la práctica, el procedimiento puede y debe abreviarse de todos modos, ya que el trabajo se realiza en función de necesidades y decisiones prácticas y no hay que examinar todos los detalles. Por ejemplo, se trata de planificar intervenciones terapéuticas, encontrar un candidato adecuado en la selección de personal, formular los siguientes pasos en la gestión de casos, etc. El uso del procedimiento es manejable, limitado y tiene lugar en un contexto orientado a los costes. Esto conlleva limitaciones naturales. Según nuestra propia experiencia, en un grupo formado por N = 5-8 personas, en su mayoría no académicas, se puede obtener un resultado muy aceptable y relevante en la práctica en 35-45 minutos. Esto es entonces suficiente para planificar acciones apropiadas al caso en el contexto respectivo. Cualquier imprecisión forma parte del proceso. Esto se compensa con el hecho de que en contextos de práctica no académicos, prevalecen preguntas limitadas que se puede contestar en un periodo de tiempo limitado.

No obstante, es posible que durante el análisis surja la sensación de que "ahora ya lo tenemos" y se pasan por alto o se descuidan las reglas del examen crítico del texto y del trabajo sobre la propia ceguera operativa. No se trata de acabar rápido o de reforzar los propios supuestos, sino se trata "simplemente" de reconstruir objetivamente el sentido y formular estructuras latentes. Se continua el análisis hasta que la hipótesis de estructura de caso que se ha encontrado se reproduce realmente en su totalidad (Oevermann, 1996a). Del mismo modo, en la comprobación crítica de hipótesis, se trabaja críticamente de hecho. Dado que la lógica interna se reproduce y desarrolla de nuevo en cada pasaje basta con examinar detalladamente algunos pasajes y obtener así una visión general del caso.

El trato con el propio impulso interior afecta tangencialmente al nivel emocional-motivacional y no al nivel puramente cognitivo. En consecuencia, en un caso crítico, es necesario examinar por qué alguien quiere acabar (más) rápido y cómo se puede manejar esta situación. En esos momentos, la supervisión externa o la supervisión entre compañeros es indispensable para no poner en peligro el análisis.

11.9.6.2 Libertad contextual

Un error típico es siempre el uso de información que no ha aparecido previamente en el texto y que, por tanto, en sentido estricto, no se puede utilizar para una interpretación. Al mismo tiempo, sin embargo, la generación de hipótesis debe estar libre de contexto. Para ello, hay que inventar historias legítimas que puedan dar a un pasaje del texto un trasfondo interpretativo plausible. Sin embargo, no se trata de interpretaciones cotidianas o de psicología o sociología de cocina que vienen a la mente ad hoc, sino de interpretaciones que conllevan consecuencias y que, por tanto, muestran efectos y permiten predicciones para contenidos textuales posteriores que deben ser comprobados.

11.9.6.3 Literalidad

No tomar los textos literalmente significa interpretar el sentido figurado – por los motivos que sean – en un texto. Así pues, la literalidad no se tiene en cuenta cuando los analizadores utilizan sus ideas preconcebidas para interpretar un pasaje del texto en lugar de utilizar realmente cada palabra realmente pronunciada como una cadena de información que se construye una sobre otra. No se trata de lo que alguien podría haber dicho o querido decir, sino de cuáles son las consecuencias si alguien dice exactamente lo que realmente aparece en el texto. En el lenguaje cotidiano, esto se expresaría como la diferencia de "Aquí no estamos en lo-que-deseamos, sino en lo-que-estamos". Por lo tanto, si la obra no se aproxima lo suficiente al texto, el contraste entre posibilidad y realidad se produce de forma insuficiente. En ese caso, se analiza un contraste "posibilidad a posibilidad", es decir, "construcción a construcción". Tal contraste se puede decidir difícilmente o incluso falsarse, si el texto real no constituya la base real de la argumentación.

11.9.6.4 Secuencialidad

La secuencialidad siempre se viola cuando no se procede a lo largo de la estructura natural del texto, sino que se salta hacia delante y hacia atrás en el texto de forma inadmisible. Aunque esto es posible o incluso útil en el paradigma de la codificación, cuando se trata de codificar o comparar pasajes de texto (véase el capítulo 9.5.6), esto está sencillamente prohibido en el análisis secuencial. En el fondo, el error consiste en no saber limitarse a la información realmente disponible en el momento actual.

11.9.6.5 Extensividad

Una ponderación inadmisible y un tratamiento no equitativo de los módulos de texto disponibles (palabras, ...), como la exageración o el énfasis inadmisible, el descuido, la omisión, la declaración de carencia de importancia, la reserva para más adelante, etc., conducen a distorsiones sistemáticas en la interpretación. Básicamente, todo tiene la misma importancia, sea lo que sea. No es tarea de los analizadores destacar o devaluar la información, sino reconstruirla.

11.9.6.6 Parsimonia

La parsimonia se viola cuando se recurre precipitadamente a explicaciones patológicas o que se apartan de la norma general sin pruebas suficientes. Nota – se necesitan razones de peso para desviarse de la normalidad. Hay una compulsión a justificar la desviación de la normalidad concretamente a través del propio texto.

11.10 Hermenéutica objetiva y teoría fundamentada

La Hermenéutica Objetiva fue creada por Oevermann y colegas como un paradigma de investigación independiente, sin abordar otros puntos de integración y conexión con los enfoques de investigación existentes. Hildenbrand (2006a), por su parte, ofrece una propuesta para combinar la Hermenéutica Objetiva y especialmente el análisis secuencial con la teoría fundamentada según Glaser y Strauss (1967). Este enfoque, que pertenece al *análisis de casos individuales*, selecciona los casos según el principio de la teoría fundamentada del *máximo contraste* (véase la Fig. 11.2), utiliza el análisis de secuencias para reconstruir la estructura del caso y utiliza el criterio de saturación teórica a través de todos los casos para determinar el número de casos en la práctica de investigación y alcanzar un nivel adecuado de información.

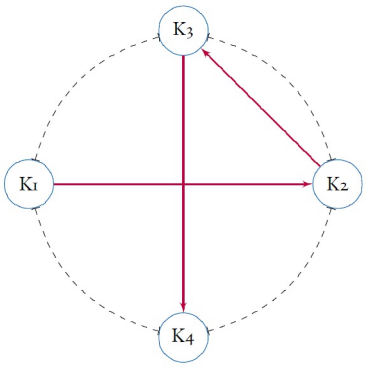


Figura 11.2. Principio de máximo contraste

Este enfoque nos parece muy razonable, ya que no implica grandes incoherencias teóricas en la práctica y ofrece un marco bueno y eficaz para explorar un nuevo campo de investigación, por ejemplo. Encontrará más detalles en Hildenbrand (1999). La hermenéutica objetiva parece haber tenido poco en cuenta esta ampliación y no la ha incluido en su canon. Al igual que el programa de investigación Teorías Subjetivas en Psicología, la Hermenéutica Objetiva parece delimitarse a sí misma en lugar de abrirse a la integración. Si uno se aplica a sí mismo los principios de la Hermenéutica Objetiva, tiene que preguntarse hasta cuándo seguirá con los hábitos y cuándo éstos fracasarán, de modo que se producirá una crisis que podría cambiar permanentemente la práctica vital de la Hermenéutica Objetiva. El mismo principio puede aplicarse a muchos enfoques comparados de la ciencia.

11.11 Integración de métodos

Como se analiza con más detalle en el capítulo 13, existen diversas formas de practicar la integración de métodos. Estructuralmente, todos estos intentos se reducen a dos tipos:

- Los datos de un paradigma se analizan con métodos de otro.
- La secuencia de "Recogida de datos – Análisis – Interpretación" de un paradigma se complementa con una secuencia comparable de otro paradigma y se vinculan los hallazgos respectivos.

Aunque sin duda puede tener sentido contar códigos y realizar análisis estadísticos (especialmente de forma gráfica o con métodos AED, véase cap. 5) dentro del paradigma de codificación, esto no es posible ni sensato con los datos generados por el análisis de secuencia. Por supuesto, se podría proporcionar una estructura de casos con códigos y, entre casos se podrían generar códigos, contarlos, etc., pero esto supondría desperdiciar el gran esfuerzo que ya se ha realizado con el análisis de secuencia. De hecho, estaríamos volviendo del Porsche al 2CV. Del mismo modo, habría que preguntarse qué se supone que hay de nuevo si ya se dispone de una estructura de casos completa y examinada empíricamente de forma crítica. A pesar de toda la defensa de la integración de métodos, la integración de métodos a este nivel de datos no sería aconsejable.

El punto dos parece mucho más interesante, a saber, un enfoque analítico secuencial y la vinculación de los resultados de dicho análisis con otras fuentes de datos, análisis e interpretaciones – pero este enfoque es muy poco frecuente. Un ejemplo es el trabajo de Studer (1998), en el que, por un lado, basándose en la estadística de Bayes para muestras pequeñas (Studer, 1996b; Bretthorst, 1993), se comprobó de forma conservadora el éxito de un centro hospitalario para adictos en Suiza. En paralelo, se dispone de casos del centro que completan el panorama general. Por un lado, esta integración de métodos ilustra la calidad y el procedimiento de la institución basado en un concepto sistémico profundo orientado a casos (Gürtler, Studer & Scholz, 2012). Por otro lado, las estadísticas examinan de forma crítica y conservadora las tasas de éxito de la institución durante un periodo más largo de tres años más un periodo catamnésico. En un estudio de

seguimiento de Gürtler, Studer y Scholz (ibíd.), los casos seleccionados así como los examinados en el estudio de Studer (1998), fueron examinados de nuevo muchos años después de abandonar la terapia de una manera secuencial-analítica sobre la base de entrevistas biográficas y se derivaron pronósticos para futuras historias clínicas. En el capítulo 5.5.7 ampliamos el análisis y los resultados del estudio de Gürtler, Studer y Scholz (2012) sobre el potencial de reintegración catamnésica. La pregunta de investigación era: "¿Qué trayectorias biográficas siguen los antiguos clientes?" y "¿Cómo se puede relacionar los hallazgos con la información registrada en la terapia y antes de ella en un contexto congruente?". A continuación, se creó una secuencia paso a paso en todos los casos sobre cómo salir de la drogadicción. Además, se estimó el nivel de recuperación alcanzado en cada caso.

Los casos se contrastaron entre sí en relación con el nivel y el potencial de su propio tratamiento de la adicción. El nivel de recuperación alcanzado se estimó para cada caso y los casos se contrastaron entre sí en términos de su nivel y potencial para hacer frente a su propia adicción. Así pues, la integración de métodos puede funcionar según el principio de bloques de construcción si existe una pregunta significativa.

11.12 Análisis secuencial asistido por ordenador con AQUAD 7

El método de análisis de secuencia analógico, es decir, con papel y lápiz, es costoso e ineficaz en la práctica. Especialmente cuando se trata de redactar un informe o una documentación, apenas es posible reproducir adecuadamente la evolución exacta de las distintas hipótesis, su comprobación y condensación en lecturas y, posteriormente, en la hipótesis de la estructura del caso. En realidad, esto sólo es posible con una grabación en vídeo o cinta magnetofónica, que habría que ver íntegramente para el informe a fin de poder ofrecer una imagen completa de los resultados, además de las grabaciones escritas. Obviamente, esto lleva mucho tiempo y no es especialmente claro.

Por este motivo, hemos llevado a cabo todo el proceso de análisis de secuencias, que consiste en

1. generación de hipótesis
2. comprobación de la hipótesis
3. posible falsación de la hipótesis de estructura de casos propuesta

implementado en el software de código abierto AQUAD 7 (Huber, 2019) (véase Fig. 11.3).

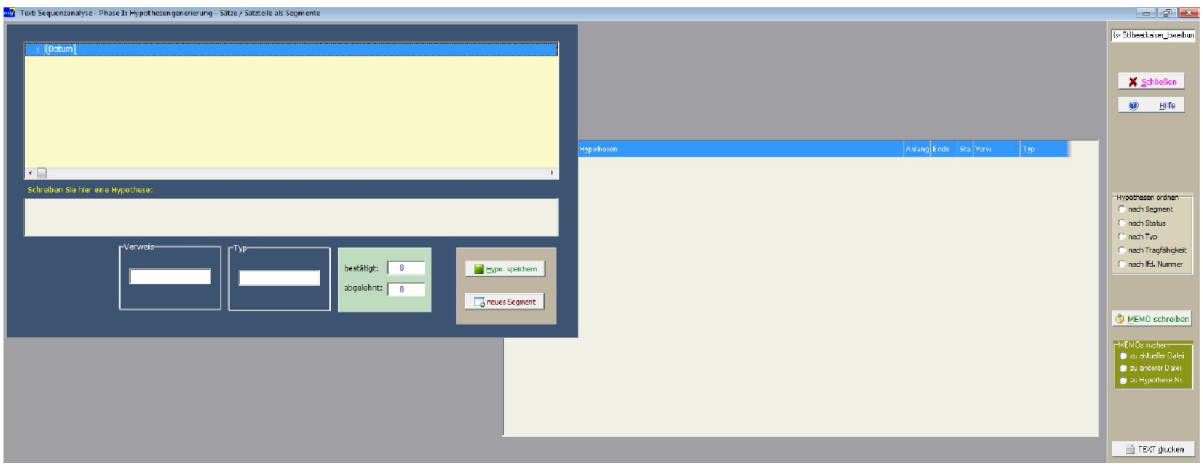


Figura 11.3 *Secuencia de trabajo analítico con AQUAD 7*

Los detalles de la implementación del software, las justificaciones y las implicaciones (por ejemplo, para la revisión por pares o la revisión de trabajos de cualificación) de este procedimiento se pueden encontrar en Gürtler y Huber (2016) y Huber y Gürtler (2012, cap. 5). La figura 11.3 muestra cómo esta parte del programa funciona y está estructurada visualmente.

El procedimiento en sí *no se ha modificado*, por lo que la aplicación informática es puramente un sistema de *gestión de hipótesis y datos*. Debido al registro completo del proceso de análisis, el propio procedimiento de análisis de secuencia se hace accesible para los más variados controles: Se puede realizar meta-análisis con respecto a los errores que puedan producirse durante el análisis o se puede obtener una visión de los procesos de análisis típicos. Se podría preguntar cuántos cambios sufre una hipótesis y cómo éstos pueden diferir cualitativamente entre sí antes de que todo conduzca a una hipótesis preconcebida de estructura de caso, o si hay momentos típicos en los que se formulan hipótesis especialmente sostenibles y en qué pasajes del texto suele ocurrir esto. Sería interesante averiguar cuántos intentos de falsación hacen los investigadores antes de decidirse por una hipótesis final de estructura de caso. Las posibilidades son casi sin límites.

Sin embargo, es mucho más importante hacer que el resultado de un análisis de secuencia sea accesible a un escrutinio externo, ya que el uso de ordenadores significa que cada hipótesis, cada examen crítico, etc., queda registrado. Esto conduce directamente a la protección de datos, que a su vez requeriría una publicación aparte. En cualquier caso, es sorprendente cómo se realiza la gestión interna de la calidad en la Hermenéutica Objetiva desde hace décadas, si el procedimiento no es accesible a un control externo tan sencillo (sin soporte informático). Esto sólo puede significar que el proceso de análisis de datos en sí mismo rara vez es objeto de investigación, sino que los resultados se discuten por su plausibilidad, pero no el camino para llegar a ellos.

11.13 Análisis de secuencia – Ejemplo de un estudio de caso

Retomamos el estudio de caso del capítulo 10.1 (análisis cuantitativo de texto) o 10.1 (paradigma de codificación) y ahora lo examinamos secuencialmente (Gürtler, Studer & Scholz, 2012, Cap. 2 para más detalles sobre el estudio de casos). En primer lugar, desarrollamos una demencia profesional y olvidamos lo que ya hemos hecho hasta ahora con el conjunto de datos. Esto es necesario porque estamos trabajando de tres formas distintas con el mismo material en este libro, y uno de los autores ya ha trabajado intensamente con el material hace años y, por tanto, lo conoce bastante bien. Descubriremos si esto es un éxito o no cuando otros comprueben nuestro análisis. Comprobar el propio análisis sería ir demasiado lejos. A veces es mejor tener una segunda y tercera perspectiva de alguien tercero, cuarto o quinto. Además, realizamos el análisis sólo para ilustrar el proceso básico de análisis y no entramos en todos los detalles, ya que esto iría más allá del ámbito. De todos modos vamos más allá del marco, ya que pensamos que tendría sentido explicar en el sentido de pensar en voz alta, qué pensamientos y argumentos surgen en el proceso de investigación y cómo tratarlas. Así que formulamos con la mayor precisión posible nuestros pensamientos y nuestras comparaciones, por qué interpretamos algo de esta manera y no de otra, dónde quedan incertidumbres respecto a una interpretación, qué parece más claro, etc. Esto sirve para el principiante del modelo de varias formas – una vez realmente para *aprender del modelo*, pero también como antimodelo. Puede ocurrir que un lector descubra que no se puede ni se debería trabajar así porque ... – entonces habrá razones para ello. Hay que captarlas y hacerlo de forma diferente y mejor a como lo hacemos posteriormente. Todo es información.

Normalmente empezaríamos con un análisis de los datos objetivos del caso. Sin embargo, nos abstraemos y nos olvidamos del sexo, la edad, la educación y la profesión, la situación familiar y ocupación, lugares y todas las demás características objetivas de la persona investigada. En su lugar, hacemos un pequeño análisis contextual de la situación en sí, es decir, reconstruimos la visión de la institución y su visión de un posible solicitante de una plaza de terapia de adicción. El punto de partida es la abstinencia, que suele tener lugar en un pabellón psiquiátrico. El retiro es el lugar desde el que se formula, escribe y envía por fax la solicitud de una plaza de terapia.

11.13.1 Análisis del contexto - solicitud de una plaza de terapia de adicción

La pregunta del análisis de contexto pegadiza es ¿qué esperamos normalmente de un cliente motivado que está en un centro psiquiátrico y escribe una carta de solicitud de una plaza en terapia de drogadicción?

11.13.1.1 Situación general de los adictos en abstinencia

En primer lugar, analizaremos el contexto en el que se encuentran los adictos cuando escriben una carta de solicitud de una plaza de terapia desde el síndrome de abstinencia. ¿Qué podemos tomar como un caso normal? Nos centramos en la situación de un adicto en abstinencia y en el aspecto de su planificación para el futuro y menos en las sensaciones físicas subjetivas actuales, pensamientos, deseos, honores y miedos y todo lo demás que conlleva. Sólo eso ya sería motivo de un largo excursus.

El punto de partida es la retirada del uso anterior – y la retirada establece el prerrequisito mínimo para un lugar en la terapia y un cambio del extremadamente desfavorable estilo de vida, ya que las personas con drogas en la sangre generalmente no consiguen una plaza de tratamiento en centros de hospitalización. Un centro de hospitalización es un asunto a largo plazo, es decir, se debe tener en cuenta al menos un año de terapia (para más detalles, véase Studer, 1998). Es que toda persona en síndrome de abstinencia ya ha llegado al final y ve la terapia como el último perfeccionamiento. Esto debería tener un efecto positivo en la motivación de la persona, ya que entonces una orientación sana hacia la realidad determina en cierta medida sus acciones. Por ejemplo, hay personas que prefieren aprender como se podría consumir drogas, pero sin sufrir sus nocivos efectos secundarios, o hay personas que, por su pareja o familia, preferirían más bien acabar en abstinencia a la fuerza, pero no tienen una motivación real para la terapia. Aunque hay solicitantes motivados, también los hay desmotivados y los que quieren utilizar la terapia para sus propios fines, pero no para conseguir una vida verdaderamente libre de drogas para sí mismos o no para trabajar en sus propios modelos mentales-somáticos.

La situación de abstinencia podría ser el comienzo de la autorregresión. Pero frente a la desintoxicación del cuerpo, las otras personas que también están presentes en este lugar y el ambiente general, se puede suponer que la desintoxicación no es un lugar de descanso, sino más bien uno de desilusión en sentido literal. Durante este tiempo el problema es averiguar hacia dónde va a ir la propia vida a partir de ahora. Esto no funciona realmente bien. Esto va acompañado de la cuestión de hasta qué punto se es consciente de uno mismo para ver que hay que cambiar. De hecho es muy poco probable que uno pueda cambiar su vida en este momento sin ayuda externa. Si realmente existe la perspicacia y la motivación para iniciar una terapia contra la adicción, el reto es elegir la institución adecuada. Sin imponer una regla, parece plausible que no todos los entornos son siempre adecuados para todas las personas en todas las circunstancias. Se trata de una afirmación neutral sobre el ajuste mutuo entre cliente e institución y atribuye ningún motivo de fracaso potencial en ninguna dirección. Una institución es consciente de ello y selecciona a los clientes en función del ajuste mutuo. A principios de los noventa, la selección de una institución aún no podía hacerse a través de Internet. Por lo tanto, la información inicial sobre las instituciones potenciales se facilita mediante comunicación oral (callejera) o a través de folletos y carteles (por ejemplo, en psiquiatría). Habría que profundizar en este aspecto, ya que no se dispone de información al respecto.

Mientras que en una solicitud de empleo se podría hacer hincapié en una buena formación, las notas, la experiencia previa, los trabajos anteriores, etc., esto es menos relevante a la hora de solicitar una plaza en terapia de adicciones. Por supuesto, se pueden citar abstinencias anteriores, terapias, etc., pero la mayoría se abstendrá de hacerlo, ya que podrían de parecer peores de lo que creen ser en su falsa autonomía. El reto en una carta de solicitud es ser honesto y franco sobre la propia situación actual y la situación y la motivación para la terapia. Por motivación pegadiza para la terapia nos referimos explícitamente sólo a la motivación inicial para la terapia, ya que cabe suponer que la terapia inicia tantos cambios que, en el mejor de los casos todos los niveles (cognición, emoción, motivación, acción, cuerpo, etc.) cambian y se desarrollan. Esto incluye motivación para la terapia, que puede y debería cambiar, por ejemplo, de la pura supervivencia hacía la educación, la salud, la asociación, el deporte, etc. Un progreso relativo creciente en la autonomía significa poner en práctica las propias ideas, decisiones y aceptar y vivir con las consecuencias, tanto positivas como negativas. En el caso de la motivación de la terapia pegadiza, suponemos que es tanto auténtica como no

auténtica. Es importante elaborar las partes respectivas para obtener una imagen completa del solicitante. No se trata de desacreditar a las personas ni de acusarlas de falta de autenticidad o incluso de mentir. Todos esos pensamientos están fuera de lugar. Más bien, es importante comprender cuál es la mejor manera de ayudar a las personas. Para ello necesitamos conocer tanto sus capacidades como sus lados más oscuros.

11.13.1.2 Expectativas típicas de una carta de solicitud

Una carta de solicitud consta de dos partes: solicitud y carta. En cuanto al contenido, esperamos que alguien nos dé buenas razones por las que deberíamos acogerle. Un proceso de solicitud habitual en el caso de un puesto vacante puede constar de muchos pasos, dependiendo de la ubicación del puesto en una empresa. No todos ellos se dan en todas las solicitudes: Anuncio o búsqueda por parte del headhunter, contacto inicial, llamada telefónica preliminar o comunicación con preguntas sobre el puesto y otras características del mismo (por ejemplo, el puesto en sentido estricto, el entorno de trabajo, la empresa, el salario, etc.) y una solicitud formal por escrito con foto, CV, carta de presentación, posiblemente una carta de motivación e ideas o conceptos para el para el puesto, referencias de formación, estudios, actividades anteriores, cartas de recomendación, lista de experiencia previa y posible información adicional sobre el candidato. El tipo de información adicional (por ejemplo, lista de publicaciones en el ámbito universitario) varía en función del tipo de puesto, de la experiencia que ya se tenga, etc. En función de la reacción positiva e interés de la empresa, seguirá una solicitud o entrevista, así como, dependiendo del puesto, sigue un procedimiento por etapas (por ejemplo, trabajo de prueba, diversas pruebas, centro de evaluación, etc.). El último paso es el rechazo o la contratación satisfactoria, de modo que el proceso finaliza con la negociación del contrato. Es decir, se trata de un proceso continúa, dependiendo del puesto de trabajo, con las negociaciones del contrato o, de forma poco espectacular, simplemente con el inicio del trabajo. Se trata, por tanto, de un proceso que puede implicar una gran variabilidad de esfuerzo. Una duración de meses hasta la contratación no es inusual. Pero no sólo en el caso de trabajos sencillos la contratación puede producirse muy rápidamente, por ejemplo cuando se necesita a alguien urgentemente. En una cafetería temporal, esperamos un proceso de solicitud muy corto, verbal más que escrito, y prácticamente un apretón de manos. En un trabajo de comerciante ya se trata de la formación existente o el interés por un aprendizaje, así que, como mínimo, un informe escolar o una referencia laboral previa. Los procesos son similares, pero hay las diferencias. Aquí son más cortos en muchos sitios, mientras que allí merece la pena que una empresa o un propietario de una empresa invierta en un proceso de solicitud más largo, porque en los puestos más altos hay más en juego. Un buen proceso, por cierto, contrata a las personas según su potencial para el futuro y no (sólo) según las actividades y aptitudes pasadas.

Sin embargo, cuando se solicita una plaza de terapia en un centro hospitalario de adicciones, no se trata de cubrir una plaza de personal, sino de un cliente que paga en principio (seguro médico, autopagador). Esto significa que el punto de partida no es que una empresa quiera emplear a alguien que tenga un empleo productivo, que genere un beneficio monetario para la empresa o que sea productivo de alguna otra manera – y que se le pague por ello. Más bien, se adquiere un cliente que aporta dinero y al menos él que soporta el coste (por ejemplo, la compañía de seguros médicos) espera algo a cambio. En el mejor de los casos, se trata de una vida posterior libre de drogas y de una reintegración en la sociedad. Así que ocupamos un puesto que nos paga – y no al revés – y lógicamente, se espera algo a cambio, a saber, una terapia exitosa con los consiguientes cambios en los niveles de cognición, motivación, acción, cuerpo e interacción social. En última instancia, tanto la institución como el cliente se aplican, pero el cliente da el primer paso y se encuentra en una situación menos poderosa.

Por un lado, tenemos que transformar nuestras expectativas hasta el estado de un drogadicto en abstinencia. Por otro lado, sí esperamos una carta formal correcta con razones personales relevantes que sugieran que debemos llevar exactamente a este cliente o al menos a invitarle a una entrevista. Somos realistas, en retirada en psiquiatría y a principios de los años 90 los equipos informáticos, la incrustación de fotos en la escritura, etc. no desempeñan prácticamente ningún papel. Las expectativas no se materializan que estándar en las cartas de solicitud hoy en día y para las que existe, en caso de duda, alguna app. En cambio, se esperan frases completas en alemán, un asunto, un saludo correcto y un saludo formal final y

datos de contacto del remitente y el destinatario. Formalmente la fecha (normalmente en la parte superior del membrete), aunque una máquina de fax imprima estos datos como un informe de fax cualificado.

En cuanto al contenido, nos contenemos y no hacemos demasiadas suposiciones, porque los planes de vida y las situaciones de las personas son muy heterogéneos. Una carta de presentación para una plaza en una clínica de adicciones debe contener ciertos elementos básicos que actúen como requisitos mínimos: *Se trata de una vida libre de drogas para los clientes*. Ese es el objetivo de la terapia. No esperamos los siguientes elementos en una formulación o secuencia exacta, pero de una forma u otra deben darse. Si faltan elementos de la siguiente lista, se trata de información importante sobre un cliente. En cuanto al contenido, esperamos afirmaciones sobre los siguientes temas:

- Una valoración razonablemente realista de la propia situación, es decir, no poder llevar una vida adecuada. La forma de expresar este hecho da una buena imagen de cómo se siente la persona, cómo le gustaría presentarse a la persona en cuestión, y dónde prevalece la falsa autonomía y dónde no.
- Aceptación de la situación de necesitar ayuda externa para una vida libre de drogas, concretamente la de un centro de terapia de adicciones. Tal afirmación es similar, en un sentido más amplio, al planteamiento de los
- grupos anónimos de autoayuda, que comienzan con un reconocimiento de la propia situación, pero luego pasan a otros puntos con los que no siempre estaríamos necesariamente de acuerdo (para más detalles sobre los grupos anónimos, véase detalles sobre los grupos anónimos, Scholz, 1992 o Studer, 1998).
- Una declaración con fundamentos sobre la propia motivación para la terapia. Esto incluye un objetivo de algún tipo en cuanto a lo que se supone que la terapia debe lograr – por ejemplo, volver a la vida normal, trabajar sobre la propia adicción, romper con el consumo de drogas y el estilo de vida asociado a la delincuencia relacionada con las drogas, o el nombramiento de algún otro motivo personal y comprensible. Una razón perfectamente legítima sería el deseo de sobrevivir. Una frase de este tipo integraría casi todos los puntos anteriores.

La expresión de la propia motivación para la terapia se corresponde con el nivel o el estado de la propia autonomía dañada y, por tanto, refleja directamente las conclusiones sobre el cliente. El significa que esperamos indicaciones claras de dónde se localiza el daño, basadas en la autonomía limitada. En este sentido, no significa que tras analizar un "fallido" carta de solicitud digamos: "Oh, no aceptaremos a esa persona, no parece honestamente motivada". Las cartas de solicitud fallidas no existen, sólo la información. En consecuencia, utilizamos un análisis para evaluar el statu quo, identificar las áreas de mejora más importantes y estimar los potenciales y recursos existentes. Todo esto junto en comparación con otras cartas de solicitud, se decide entonces sobre las entrevistas. Las plazas de terapia también se cubren en función de la ocupación. Así que hay varios factores que en parte no tienen nada que ver con los clientes potenciales, pero que deciden sobre un lugar de terapia. Sin embargo, los clientes difícilmente serán conscientes de ello cuando redacten su carta de solicitud. Cabe suponer que se ven a sí mismos ante todo.

Lo que dejamos fuera del caso aquí son los clientes de las medidas. Por ejemplo, a principios de los años 90, era habitual en Zúrich, en la institución "start again" aquí estudiada, que los clientes recibieran *terapia en lugar de castigo*. Así, estos clientes acababan en prisión por orden judicial. Pero en lugar de estar en prisión se les dio la oportunidad de trabajar y cambiar fundamentalmente su adicción y, por tanto, así cambiar fundamentalmente su conducta delictiva. Para más detalles, véase el estudio de evaluación de Studer (1998), que fue financiado por la Oficina Federal de Justicia Suiza (BAJ) para evaluar el enfoque sistémico profundo de la institución "start again". El autor llega a la conclusión que los clientes se benefician de la terapia contra la adicción tanto como cualquier otra persona, a condición de que el tiempo sea esencial. Esto significa que cuanto más larga sea la estancia, más confianza puede ganarse la confianza de estos clientes y pueden desarrollar su potencial como todos los demás clientes.

En resumen y basándonos en los patrones de comportamiento típicos en el contexto de la adicción (Scholz, 1992; Studer, 1998) y en vista de las condiciones situacionales en el síndrome de abstinencia, formulamos la expectativa de que una carta de solicitud se caracterice por los siguientes aspectos formales y externos componentes externos que van más allá de las consideraciones de contenido ya arriba:

- El texto es relativamente breve, porque no hay mucho que decir, salvo el deseo de una entrevista. Si una carta de solicitud de este tipo es inesperadamente larga, se trata de información propia y no corresponde a las expectativas. Esto puede entonces tener posiblemente un significado y debe examinarse con más detenimiento. Sin embargo, tampoco debe tratarse de una sola frase. Se podría decir que cuatro a ocho líneas de fax pueden ser suficientes, escritas a mano y dependiendo del tamaño de la letra, ya que el equipamiento informático de la época era mínimo y no se tenía necesariamente acceso a una máquina de escribir.
- Mucho texto superfluo o explicaciones que no están directamente relacionadas con la motivación de la terapia en sentido estricto y el objetivo del lugar de terapia, explícitamente no las esperamos. Las desviaciones de esto requieren un análisis detallado.
- Los componentes formales de remitente, destinatario, persona a la que se dirige, saludo, texto y conclusión con fórmula de saludo y fecha.

Básicamente, esperamos una carta de solicitud normal como para una oferta de empleo, pero redactada a un nivel representativo de la clientela pertinente. Currículum vítae, trabajos de prueba, experiencia previa y añadidos similares no se esperan. Traducido al contexto de adicción, podrían ser: retiradas anteriores, tabuladas y enumeradas ordenadamente de forma cronológica. Nada de esto está dentro de nuestro horizonte de expectativas, pero si lo está, tal vez sea una información importante. Cabe suponer que la propia redacción implica un gran esfuerzo. Por el aspecto exterior sabemos que se trata de un fax, un método habitual de registro de la comunicación en esta época. En consecuencia no esperamos una carta, papel de carta limpio y sobre, etc., ya que esto no corresponde a los medios de comunicación elegidos a disposición de los clientes potenciales en retirada. En función del certificado de escolaridad y del nivel de estudios, no nos atrevemos a interpretar las faltas de ortografía. Éstas pueden deberse simplemente a la falta de educación.

11.13.2 La carta de solicitud en detalle

11.13.2.1 La unidad de análisis

Elegimos una unidad muy pequeña como unidad de análisis. Esto significa que sólo analizamos frases parciales, leyendo el texto únicamente hasta el final de la frase parcial, y a partir de ahí ni una palabra ni un signo de puntuación. Esto requiere una cierta ceguera profesional unida a una buena dosis de pedantería y obsesión. A estas alturas ya nos atrevemos a hacerlo. Podríamos hacer el análisis en AQUAD 7 (Gürtler & Huber, 2016). Pero por razones analógicas y debido a un cierto entusiasmo por el papel y el lápiz y como homenaje al trabajo tradicional en la Hermenéutica Objetiva, simplemente gestionamos nuestro análisis sin el ordenador. Sin embargo, si lo desea, puede reproducir cada paso con el ordenador. Las instrucciones para ello se encuentran en el manual de AQUAD 7.

Trabajamos de tal manera que imprimimos el texto (ver fig. 11.4) después de estructurarlo en partes de frases (véase fig. 11.5). A continuación, la hoja se dobla de tal forma que sólo podamos leer la sección del texto lista para el análisis y nada más. A continuación, el análisis sigue los principios del análisis secuencial (véase cap. 11.9).

11.13.2.2 Análisis secuencial del texto

Comenzamos con las formalidades y esperamos una fecha, un remitente completo con nombre completo y dirección de contacto y lo mismo para el destinatario, la institución de terapia de adicción. Dado que se trata de formalidades, todos los clientes deberían poder hacerlo. Las desviaciones de la norma tendrían que discutirse intensamente.

```
[Fecha]
```

La fecha corresponde a la expectativa. Como tradicionalmente el remitente figura antes que el destinatario, ahora esperamos la dirección de contacto del cliente potencial, seguida de la del destinatario.

```
[Nombre del cliente]
```

Esto corresponde a las expectativas. Se trata de un cliente masculino llamado Beat Kaiser (nombre completamente anónimo, véase Studer, 1998).

```
Actualmente [nombre psiquiatría]
```

Esto también coincide con las hipótesis preliminares. Dado que la persona en cuestión se encuentra en estado de abstinencia en psiquiatría lo que descarta una salida. Esto entra exactamente dentro de las expectativas de que ninguna dirección privada posiblemente existente figure como contacto. Lo ideal sería que la transición de del síndrome de abstinencia a la terapia, para que todo parezca coherente hasta ahora.

```
Start Again
```

Esto también está dentro de lo esperado.

```
Glärnischstr. 157
```

Nada nuevo.

```
8708 Männedorf
```

Ahora la fecha, el remitente y el destinatario figuran en su totalidad. Ahora esperamos una línea de cubierta formal dirigida al lugar de la terapia. Porque de eso se trata.

```
Solicitud de entrevista
```

```
{   1} [Datum]
{   2} [Name Klient/in]
{   3} z.Zt. [Name Psychiatrie]
{   4}
{   5}
{   6}
{   7} Start Again
{   8} Glärnischstr. 157
{   9} 8708 Männedorf
{  10}
{  11}
{  12}
{  13} Bewerbung um ein Vorstellungsgespräch
{  14}
{  15}
{  16}
{  17} Hoi zäme!
{  18}
{  19} Ich habe Euer Konzept gelesen und könnte mir vorstellen, dass ich bei
{  20} Euch gut aufgehoben wäre. Es kostete mich einige Zeit, mir eine Zukunfts-
{  21} perspektive zu entwickeln. Diese würde ich Euch gerne in einem persön-
{  22} lichen Gespräch erläutern.
{  23} Ich hoffe, dass Ihr mir diese Chance trotz meinem relativ hohen Alter
{  24} gewährt.
{  25}
{  26}
{  27}
{  28} Mit freundlichen Grüssen,
{  29} [Unterschrift Name Klient/in]
{  30}
```

Figura 11.4. *Carta de solicitud – análisis de secuencia (esquema del material textual)*

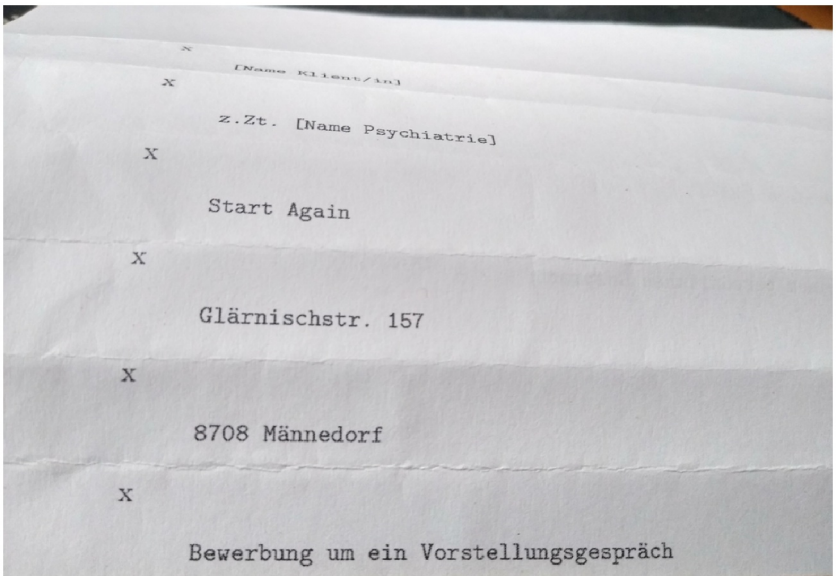


Figura 11.5. *Carta de solicitud – análisis de secuencia (foto del procedimiento en la práctica)*

En principio, esto se corresponde con lo esperado en términos de dirección. Sin embargo, llama la atención que se trate de entrevista y no de una plaza en terapia, el objetivo real. De este modo, el escritor se queda corto respecto a sus propias posibilidades y evita el verdadero tema. Algo impide al escritor llegar al verdadero tema, que es la solicitud de una plaza de terapia de adicción en régimen de internado debido a una grave adicción a las drogas. Si lo comparamos con la solicitud de un puesto de trabajo, allí nadie solicita una entrevista, sino que (frase hacia el final de una carta de solicitud) "esperaría con impaciencia una invitación a una entrevista personal", ya que es el siguiente paso lógico del proceso. Aquí parece como si la entrevista fuera ya el último paso y no una parada necesaria. Así aprendemos que el escritor

- o bien quiere evitar la terapia
- o no confía en sí mismo con la terapia
- o no confía en si puede conseguir un lugar de terapia (por la razón que sea).
- o puede conseguir uno.

La situación es como cuando muchos solicitantes están sentados en una sala y uno tras otro se les pide una entrevista que han solicitado, pero sin tener un objetivo detrás o a través de la entrevista. Y después todos se van. Eso no funciona, porque o bien la selección tiene lugar antes de la entrevista según otros criterios -- y aquí no hay indicios de ello -- o todos los invitados reciben la entrevista, lo que hace absurda una solicitud. Entonces el fax sería más bien una solicitud, pero no una solicitud real. Y la entrevista es sólo una instrucción y no un fin en sí misma, ni siquiera el objetivo real. En cualquier caso, la entrevista y el objetivo se alcanzaría, y esta es la situación que nos plantea el cliente potencial.

Nos preguntamos entonces si, en el siguiente saludo y en la introducción del texto, se trata de expresarse *sometiendo, evitando o coherente con las expectativas comunes*. Someter abogaría por una mayor moderación de la necesaria. Paralelamente está la hipótesis de la evitación. Para ello, el texto debe tratar sobre los temas que están relacionados con el lugar de la terapia, pero no debe abordarse directamente ni en absoluto. Coherente con las expectativas comunes, por otro lado, sería un típico "Estimado señor o señora", o seguiría un saludo comparable. En la carta, esto va seguido de una exclamación `¡Hoi zäme!`

Este saludo es sorprendente. En primer lugar, se utiliza el alemán suizo, es decir, el dialecto local y no el alto alemán (= lengua oficial en Suiza) y luego el saludo termina con un signo de exclamación y no con una coma. Esto parece fuerte y separa el saludo del comienzo del texto siguiente. Cuando se vive en Suiza, como extranjero hay un paso intermedio en el que ya no te hablan en alto alemán, sino en el dialecto local.

Aunque, o precisamente porque el alto alemán es la lengua oficial, el acercamiento y, por tanto, la integración parcial consiste en que ya no le pregunten si uno entiende o no el dialecto local, sino que simplemente se le habla de tal manera y se da por supuesto que uno lo entiende sin más preguntas. Si esta integración parcial aún no existe, se le preguntará si entiende el dialecto local o se le hablarán en alto alemán. A pesar de toda la cortesía y amabilidad de los suizos, ambos proporciona una distancia social claramente perceptible. El uso del dialecto local por parte de un suizo significa una forma de acercamiento o cercanía social y, por lo tanto, un cambio muy discreto del rol de la comunicación basada en roles a una comunicación más difusa, como es posible entre iguales o amigos y es común dentro de la familia. La expresión "¡Hoi zäme!" se puede traducir al alto alemán como "¡Hola juntos!". El cliente no se dirige a la institución singular, sino de forma inespecífica a todas las personas que allí se encuentran en plural, pero sin referirse a un papel (por ejemplo, la dirección, el terapeuta, la limpiadora, el cuidador, la secretaria, etc.). Sin embargo, no todos estos son la persona de contacto adecuada para un lugar de terapia. Alguna persona lo hará, sin embargo... El conserje podrá recibir la carta en el fax, pero ¿también decidirá y responderá a ella? Probablemente no y eso seguramente está claro para el escritor, a pesar de la situación en retirada. Por lo tanto, lo que tiene lugar aquí es el intento entre personas que en principio no se conocen, que de hecho no existe en este momento, y aunque se supone que todos son ciudadanos suizos. Esto resulta inapropiado y, en cierto modo, demasiado grandioso y distante. Se abandona la distancia formal-profesional de fecha, remitente, destinatario y asunto y el escritor prácticamente aporrea la puerta como si *no pudiera soportar su situación* o *quisiera llamar la atención*. El signo de exclamación suena como un trombón para llamar a todos. Una hipótesis sería que *lo hace intencionadamente* o *no es intencionado y simplemente ocurre*. Si es intencionado, el escritor puede retirar inmediatamente este estilo y utilizarlo cuando sea útil desde su punto de vista. Entonces un instrumento flexible, un captador de miradas, para atraer la atención y posiblemente al mismo tiempo para distraer la atención de algo, como uno mismo. Si el escritor no puede controlar esto, la comunicación tendría que continuar en este estilo difuso, pero sería coherente con esta forma de dirigirse. A partir de estas hipótesis podemos deducir otras hipótesis sobre el comportamiento flexible y adecuado a la situación del escritor e, indirectamente, cuál es su nivel intelectual. *El uso específico* de esta exclamación hablaría de la capacidad de actuar estratégicamente en un nivel más abstracto y, por tanto, de la presencia de valiosos recursos cognitivos.

El uso del coloquial "hola" apoya la suposición de una *inespecificidad y falta de concreción* en la comunicación. De este modo se podría entrar en una casa supuestamente vacía y decir "hola" o gritar un "hola" a alguien que está al otro lado de la calle y acaba de perder algo, o a quien se conoce casualmente pero no sabe su nombre. En términos de significado de la palabra, "hola" es, por lo tanto, algo fuerte, bastante alegre, a menudo en una confusión ruidosa y cercana a los gritos. Un sinónimo de "hola" sería ulular. Se trata de llamar la atención de los demás, es decir, de personas conocidas y desconocidas. Asimismo, "hola" es un saludo, pero de carácter no comprometedor, que se utiliza, por ejemplo, por teléfono. Entonces, es coherente que "hola" sea un saludo. El *cambio al dialecto local* acentúa la figura del saludo sin compromiso e intenta crear proximidad social, pero sigue siendo inespecífico.

La cuestión ahora es si el escritor irrumpe en el texto con su motivación terapéutica, habla de su situación actual en el psiquiátrico y de su propio consumo anterior o ¿se distrae de sí mismo? A la vista del asunto y de la forma de la dirección, cabe suponer que *seguirá distrayéndose* de algún modo. Eso depende de si mantiene el estilo elegido. Ignorar la distancia profesional como en el saludo se hace bien porque *no le interesa*, porque lo *considera innecesario* o porque él quiere presentarse como *alguien distinto de lo que realmente es*. En vista de la tesis del principio, hay muchos indicios de que el cliente potencial *se preocupa por aparentar ser algo que no es*, para tender a protegerse o, al menos, a ocultarse. Entonces el lema sería ocultarse por llamar la atención. Así, a pesar del uso del dialecto local, el discurso podría ser muy concreto. Se puede considerar los elementos de dialecto y concreción independientemente el uno del otro. Dialecto y concreción seguirían siendo entonces inapropiados pero no evitables. Pero seguramente no todo el mundo responde a esta carta al mismo tiempo, ¿o existe la esperanza en el fondo de que alguien en la institución haga ya una declaración positiva sobre el escritor? Así pues, plantea posiblemente al escritor *un problema claramente definido en el plano de la interacción social* de los encuentros directos entre personas concretas, donde uno no puede esconderse tan fácilmente. Una tesis sobre la ocultación – esconderse socialmente ocurre cuando uno ha tenido *malas experiencias en el pasado* y no ha sido capaz de asumirlas, no fue capaz de defenderse adecuadamente en el momento de la experiencia y, por tanto, asume que las experiencias

futuras pueden ser de la misma calidad. Entonces uno se protege por precaución. Pero la protección puede ser así agresivamente hacia delante y exige o, en el otro caso extremo, se acorta y desaparece como de la ecuación – detrás de las palabras, menos detrás de los hechos concretos, aunque eso también es posible. ¿Continúa el texto de un modo u otro? ¿Se vuelve el texto ahora exigente *con respecto al lugar de terapia* o ¿no *hay nada en absoluto y el texto se anda por las ramas*?

He leído vuestro concepto.

El escritor sólo vuelve parcialmente al nivel formal. Por un lado, sigue la institución, se dirige a ella como a una entidad y, por tanto, sigue evitando el contacto directo para mantener al mismo tiempo la supuesta cercanía. Por otro lado, vuelve al alemán y no sigue avanzando hacia el dialecto local. Construimos así la hipótesis del *problema de la interacción social directa* y suponemos que *la dirección se compuso intencionadamente en dialecto* y no ocurrió simplemente en la exuberancia de los sentimientos. Cognitivamente, entonces, el escritor puede *cambiar bien entre niveles*, donde el nivel de proximidad social, que es de naturaleza más emocional, se le escapa un poco de las manos. Aquí sospechamos que *tiene menos control*. Esto se ve reforzado por el contenido textual, a saber, que el escritor se refiere al concepto de la institución, que ha leído y no dice nada sobre su situación actual. De este modo, se presenta como un cliente potencial diligente que ya ha realizado algún trabajo preliminar (véanse más arriba las expectativas generales de una carta de solicitud). Por otro lado, evita mencionar su situación actual en el psiquiátrico, que sin duda es emocional y físicamente muy desagradable. Una frase sobre "conceptos de lectura" podría utilizarse fácilmente en un contexto completamente distinto, por ejemplo intelectual, cuando en realidad se trata del intercambio de opiniones, argumentos y puntos de vista de la gente, por ejemplo en ciencia: "He leído su concepto..." y luego "... y me gustaría intercambiar opiniones contigo al respecto" o "... y en el proceso me he dado cuenta de ...", "... y (no) me gusta porque...", etc. El escritor *desplaza su situación real al ámbito abstracto-cognitivo*, donde se puede hablar con seguridad de conceptos en lugar de la renuncia real alejándose del consumo. Allí, en la abstracción, uno no se ve obligado a poner en práctica estos conceptos y vivirlos. Todo permanece en el ámbito cognitivo, desvinculado de la realidad experiencial y por lo tanto ya no es peligroso, pero no cambia nada en la propia vida. Los conceptos sólo se vuelven peligrosos a través de su vivencia y a través de las emociones que los acompañan. Los propios conceptos llevan una vida desapegada mientras no se ponen en práctica. Si son buenos o malos, si funcionan o no, no se puede examinar críticamente en el nivel de la abstracción. Hablar de conceptos en este contexto es *una forma de estrategia de evitación o inmunización*.

Añadimos la suposición de que la primera frase de una carta de este tipo establece la dirección fundamental de la carta y, por lo tanto, no sólo es decisiva, sino que requiere un esfuerzo especial en la redacción. Tal visión del comienzo como efecto de primacía enfatiza aún más el carácter cognitivo. Por lo tanto, es significativo si el escritor en la siguiente parte del texto permanece en lo cognitivo o pasa a lo emocional, motivacional, físico, o simplemente al nivel de la acción o la experiencia. Esto indicaría que *el elemento cognitivo está fuertemente presente, pero no es dominante*, y que las emociones, etc. están en un equilibrio más o menos fuerte. Si sigue cognitivamente, esto sugiere que las cogniciones dominan y las emociones, etc. están escindidas de ellas o, en el mejor de los casos, desempeñan un papel débil. Sin embargo, las cogniciones asumen entonces tareas que no les corresponden y que no corresponden a su calidad.

La parte de la frase está escrita en primera persona. De este modo, el escritor se presenta como actuando activamente – pero sólo en el plano cognitivo -- en el centro de la acción,, aunque desde la situación real se encuentre en un papel pasivo, de espera, proverbialmente dependiente. Se intenta así *una pseudoautonomía* que, leyendo el concepto de la institución no tiene contrapartida en la realidad. De este modo, prácticamente se derrumba antes de que pueda surtir efecto. En definitiva, este comienzo demuestra *una cierta superioridad intelectual* y y el cliente potencial posiblemente intenta ponerse *al mismo nivel que la institución*, con el fin de hablar con ella de igual a igual sobre su concepto y su adaptación a él (?). Pero entonces, ¿responderá realmente al concepto como tal, es decir, a los contenidos? Si lo hiciera, sería serio. Si no lo hace, falta otra vez la ya mencionada contrapartida en la acción real.

Este estilo se ve reforzado por el *tiempo perfecto*, que sugiere que el escritor ya ha hecho su trabajo preparatorio en el pasado, pero al mismo tiempo sigue teniendo una referencia al aquí-y-ahora, de lo contrario sólo podría haber elegido el pasado gramatical del tiempo pasado. Si hubiera elegido "Leí su concepto", la acción se habría completado y estaría completamente en el pasado. Por lo tanto, permanece en el limbo, ya que el "he" está formulado en el tiempo perfecto, aunque la lectura del concepto en sí se completó en el pasado. Si hubiera utilizado el pretérito perfecto, esto implicaría un tiempo anterior al pasado que, sin embargo, rara vez se da en el lenguaje cotidiano. El tiempo perfecto es muy común.

No queda claro cuánto tiempo hace de este pasado en relación con el presente. ¿Él leyó el concepto esta tarde y luego envió el fax, o hace unos días, o hace muchas semanas, o incluso más tiempo? No lo sabemos.

El mantenimiento del estilo de destinatario inespecífico puede, por un lado, *interpretarse como una incertidumbre*, que el redactor no sabe nada de la estructura de la institución y, por lo tanto, a quién debe dirigirse exactamente. Por otro lado, como ya se ha mencionado, puede ser una *expresión de evitación*, pero no por desconocimiento, que podría abordarse meta-comunicativamente, sino para no ir al encuentro directo. El encuentro concreto es en principio capaz de fracasar, ya que en el encuentro con personas reales puede fallar la comunicación. La comunicación puede fallar, lo que significaría la pérdida de un posible lugar en la terapia. En los próximos pasajes examinaremos repetidamente esta discrepancia, es decir, si el estilo se concreta para poder fallar en principio o no. Leer un concepto no es realmente capaz de fracasar, ya que la verdadera cuestión es si el concepto se ha comprendido realmente o si se ha intentado ponerlo en práctica, aplicarlo, tener éxito con él, etc., pero eso no se plantea en absoluto. Así las cosas, no irá seguida de *ningún contenido emocional* con referencia en *primera persona*. Siguiendo con el "hola" del saludo, sin embargo, podría ser que *la responsabilidad* por la propia emocionalidad, presumiblemente *débilmente apoyada*, *se desplace al exterior*, pero de una manera inespecífica.

Dado que el estilo de tutear se mantiene en la continuación del saludo, suponemos que lo mantendrá de forma coherente durante el resto de la carta. Por tanto, no esperamos una introducción o restablecimiento de la distancia en el sentido del papel. Lo que será interesante más adelante será el saludo del final. ¿Es un -¡ya! - un inesperado "Atentamente saludos", o un sonoro y dialectal "Tschüß!" (adiós) equivalente al saludo, o una conclusión comparable que actuaría como un paréntesis coherente al "¡Hoi zäme!"?

Así la pregunta es si el escritor cambia a la perspectiva en primera persona y continúa con un "Tu"-concepto: "Me ha inspirado, creo que puede ayudarme" o "creo que me conviene" o se queda en el nivel cognitivo y posiblemente pasa a la discusión? Porque se suelen discutir los conceptos sólo mucho más tarde, y si acaso, se les prueban empíricamente. Pero esto requiere una operacionalización, es decir, un esfuerzo considerable. Desde un punto de vista científico, no todos los conceptos pueden ser tan sencillamente si tiene éxito o no. Es difícil definir criterios claros de éxito, difícil cuando sólo se trata de conceptos. Comprobar objetivos claros es mucho más fácil. En el caso de la adicción, se puede formular un criterio de éxito de forma relativamente sencilla (Studer, 1998, capítulo 10.1). Formulados de forma conservadora, los criterios son: no consumo de drogas duras, idealmente también ningún consumo de otras drogas, ningún contacto con la policía, ningún empleo o formación, educación, contacto con amigos y/o familia o pareja, actividades físicas y aficiones, así como un trabajo continuo sobre la propia adicción. El esfuerzo necesario para poner a prueba los conceptos es muy elevado, si es que es posible en casos individuales. Para un adicto en esta fase de la vida, la discusión de conceptos es claramente *secundaria*, ya que se trata principalmente de sobrevivir y lidiar con la propia adicción y no de los posibles fundamentos científicos sólidos. Hay mucho que discutir más adelante. Esto sugiere cada vez más *discrepancia en la expresión y orientación de la cognición y la emoción* como punto central de la estructura del caso del escritor. Esto se materializa a través de la *evitación*, la *falta de concreción en la comunicación* y *la pseudoautonomía* de situarse al mismo nivel como la institución misma. Por otra parte, no vemos motivos para dudar de que el escritor haya leído realmente el concepto de la institución. Esta parte de la frase es sencilla y clara y no hay razón para creer que el escritor haya mentido al respecto. ¿Por qué debería el cliente potencial hacer el esfuerzo cognitivo de empezar una carta de solicitud sobre un concepto que no ha leído? Tal esfuerzo sugeriría igualmente recursos cognitivos. Pero no lo suponemos. Sin embargo, los argumentos apoyan la opinión de la en el escritor.

¿Qué significa en realidad una emocionalidad débil y no bien fundamentada? Aquí no se trata del exceso como sobreabundancia de emocionalidad en todas direcciones y sin límites, sino más bien de la capacidad de soportar agotadoras y estresantes experiencias emocionales, que se pueden referir a acontecimientos

traumáticos y al propio cuerpo, por ejemplo, o a difíciles experiencias sociales. Débil se refiere a la falta de estabilidad para confiar en las propias emociones y dejar que guíen nuestras acciones, y para poder soportarlas a nivel de empatía durante este tiempo de experiencia. Neurobiológicamente nuestras emociones guían y controlan nuestras acciones en gran medida, y eso está fuertemente limitado en este caso. Imaginemos el caso contrario, que una persona está equipada emocionalmente fuertemente, pero cognitivamente sólo modestamente. Entonces él o ella sería capaz de y tomar decisiones basadas en sus emociones, pero no estaría en condiciones de tomar y aplicar decisiones realmente buenas y trascendentales, porque el procesamiento cognitivo de la información está insuficientemente desarrollado. Las emociones recurrirían entonces a un conjunto muy limitado de posibilidades, que sería demasiado escaso. Las capacidades de planificar acciones, de pensar estratégicamente, deducir la evolución futura, etc., se verían perturbadas. En el caso de una débil emocionalidad y un nivel cognitivo alto, las actividades cognitivas estarían todas presentes, pero debido a la emocionalidad no apoyada, todos los planes y estrategias no se podrían llevar a cabo, ya que requieren una cierta tolerancia al estrés y una dirección clara de la acción. Necesitamos todo para poner en práctica cualquier cosa en la vida: las cogniciones, las emociones y la motivación y el cuerpo trabajan juntos. Algunos de estos factores por sí solos no son suficientes.

Aquí, en la carta de solicitud, sospechamos una *emocionalidad muy vulnerable o retraída*, superpuesta por el intelecto y que, por tanto, no puede pasar a primer plano. Esto no significa que esta persona no tenga emociones, sino todo lo contrario. Pero sí significa que el pensamiento y el sentimiento no apuntan en la misma dirección y que esto puede cambiar drásticamente. Por lo tanto, la estabilidad no se da. En casos extremos, esto significa desintegración, que las emociones y las cogniciones apuntan en direcciones diferentes.

Hacia donde lo dirijamos: Para el statu quo en este punto de la carta gira en torno a la discrepancia entre una fuerte cognición y habilidades potencialmente estratégicas por un lado y, por otro, la evitación emocional, el desvanecimiento de la realidad del ego y el desplazamiento hacia la abstracción como la de los conceptos. Así que *si el escritor quiere contradecir esta figura* deben venir los siguientes fragmentos de frases, que describen la emocionalidad desde la perspectiva del yo, la propia realidad actual en abstinencia, los inconvenientes masivos que la acompañan y la motivación de la terapia en dirección a la realidad, o al menos abordarlos o expresarlos. La auto-reflexión sin las consideraciones conceptuales también serían elementos que apoyarían invalidar nuestra hipótesis previa de *fuerte cognición con al mismo tiempo emocionalidad débilmente apoyada*, y *presencia de estrategias de evitación*. ¿A dónde vamos a partir de aquí?

Me lo podría imaginar,

Continúa como la primera parte de la frase – cognitiva y ahora después de la perfecta incluso en el subjuntivo. Esto extrema el elemento cognitivo. Faltan los componentes motivacional, emocional y de acción. El *subjuntivo subraya la actitud pasiva,* reticente, presumiblemente actitud evasiva del escritor, que no adopta una postura, sino que mantiene todo sin especificar en el limbo - ¿para no poder fallar? Cuando uno ha leído un concepto que trata sobre cosas concretas, a saber, la terapia de la adicción, entonces lo ha entendido o no lo ha o sólo parcialmente. *Si no lo ha entendido*, debe haber algo aquí que diga que es incomprensible. Pero entonces, ¿para qué solicitarla? Eso parece poco probable. Independientemente de esto, sin embargo, a medida que la comprensión de uno crece, uno desarrollará una actitud hacia ella. Esta actitud llevará a un juicio, a saber, que el concepto representa un auténtico enfoque terapéutico que uno puede o no querer probar. El hecho de la carta de solicitud ya *habla en favor de una decisión positiva*, de que el concepto se experimenta como adecuado para la propia persona. Aunque se escribieran varias cartas de solicitud no contradice el hecho de que también se considere esta institución y su concepto también como una posibilidad. Pero esto lleva a acciones reales y uno podría continuar el acercamiento con "....y quiero probarlo" o "... y quiero saber más sobre ello para ver si es algo para mí". Aquí, sin embargo, estamos hablando de una idea en modo subjuntivo, que está pensando dos veces a la vuelta de la esquina. Esta evasión de una afirmación clara contradice el hecho de que el editor escribiera la carta en primer lugar. El escritor bien podría haber elegido otra institución. Tal vez lo hizo, pero al menos en el momento de escribir la carta, aún no había recibido el compromiso de una plaza en terapia. Posiblemente razones, pero no

podemos decir nada ad hoc sobre ellas basándonos en la información del texto. En la época en cuestión, a principios de los años 90, había un gran número de centros de terapia de la adicción en el cantón de Zúrich debido al problema del barrio Plattspitz, muchos de los cuales posteriormente no pudieron sobrevivir económicamente, pero eso no es aplicable al momento de esta carta de solicitud. Entonces todavía había una amplia oferta. Esto habla a favor del hecho de que o bien se enviaron *varias cartas sin respuesta positiva*, o bien se e*ligió precisamente esta institución*. Lo primero sugiere una estrategia de solicitar ampliamente y algo saldrá bien. La otra estrategia habla de una selección específica basada realmente en el concepto o en otra información que desconocemos. No obstante, hay discrepancia entre hablar de conceptos y la realidad de la propia adicción en abstinencia.

La *orientación cognitiva* que se le imputa al escritor habla de que ha leído varios folletos y conceptos y los ha entendido. Sin embargo, también se ha decidido especialmente por este centro. No comunica esta decisión en la carta, sino que hace algo así como una retirada, al tiempo que intenta correr hacia delante. Así es como se podría interpretar en este punto una carta de solicitud, que se supone que señala un interés, pero éste se desplaza a lo imaginario y al subjuntivo y a la propia pasividad. Esto tiene algo de "me gustaría trabajar para usted ... – pero sin mí" y, en apariencia, indica que no hay que abordar las cosas directamente, sino más bien pensarlas un poco más: "Lávame, pero no me mojes". Lo interesante es la pregunta del porqué, que suponemos conducirá rápidamente al problema central del cliente.

El escritor podría abordar directamente su propia incertidumbre, pero entonces esta parte de la frase empezaría con "pero" o una palabra comparable. Sin embargo, elige el nexo "y" y evita implicarse directamente.

La omisión del sujeto "yo" no debe interpretarse de forma exagerada, ya que esta forma elíptica de la frase es común y aceptada y se utiliza a menudo para dar a las frases un uso un poco más redondo del lenguaje. Por supuesto, se podría leer "y podría imaginar", en cuyo caso el escritor se situaría dos veces en el centro y al menos una vez en el proceso como agente que asume la responsabilidad de sus propias ideas. En vista de la prevalencia de este tipo de frases, sin embargo, una sobreinterpretación de este hecho no estaría justificada en nuestra opinión.

¿Cuál es el siguiente paso? ¿Quiere imaginarse hablando sobre el concepto de una forma muy concreta en el curso posterior o hay ahora una evaluación cognitiva del concepto? Suponemos que *no va seguida de una evaluación emocional* del concepto. Una forma de frase común continuaría con "que". Esto expresaría un proceso y un desarrollo mínimo del contenido, e independientemente de si el "que" va seguido de "que su concepto funciona" o "que su concepto me conviene" o algo totalmente distinto. También podría continuar con un "deberíamos quedar para hablar". Esto último serviría para que *negara por completo su propia situación actual*, bastante difícil, y al mismo tiempo se ofreciera a hablar. Se trataría de una figura interesante, ya que el nivel cambiaría entonces de lo cognitivo a la acción y entra en la posibilidad del fracaso. Sin embargo, dado que tenemos la impresión de que con *el fracaso real y la aceptación del mismo no será realmente el centro de atención* en esta carta, pensamos que una de las primeras figuras es más probable. Así que esperamos que sea una e*xtensión o ampliación de la figura del subjuntivo cognitivo* con los mismos medios. Una hipótesis más atrevida además sería *presentar como no actuante*, aunque al mismo tiempo, en el plano cognitivo y en combinación con el saludo distanciador, surge una *pseudo-autonomía*. Esto significa que el curso posterior de la frase vuelva a *situar al "yo" en una posición pasiva* y permanezca inespecífico en lo que respecta a la evaluación subjetiva de la adecuación entre la institución y su propia persona. *No esperamos* que el escritor pase a un (abreviado) "Estoy muy mal, no puedo deshacerme de mi adicción, necesito su ayuda externa y una que me ayude a cambiar realmente mi vida yo mismo - y lo que he leído sobre su concepto me da esperanza y motivación".

Lo que debemos mencionar de nuevo es la discrepancia entre el saludo y la introducción al texto de la carta. *Ahora* consideramos el saludo como una *provocación deliberada* y una *llamada de atención*, porque el curso posterior de la carta no sigue en absoluto esta figura, salvo por la forma de dirigirse. La forma de dirigirse *puede ser habitual*, ya que es posible que el escritor ya no esté acostumbrado al trato formal con las personas – así que nos contenemos con respecto a una sobreinterpretación de la distanciación. Por el contrario, no se plantean exigencias distantes ni se embellece la propia persona como superior, lo que indicaría una pseudo-autonomía sin límites. Más bien, el escritor pasa de "irrumpir en la casa" a un papel pasivo cognitivo-conjuntivo de imaginar conceptos. De forma un tanto atrevida, podríamos interpretar ahora

el discurso como una *especie de grito de socorro*. Es decir, el escritor puede tener una larga historia de ser pasado por alto, no ser tomado en serio o ser excluido y señalado. Esto podría encajar con la supuesta emocionalidad débil y explicarla en parte. Esto hace que sea interesante cómo se comporta la persona en grupos. Con una e*mocionalidad débil*, según las hipótesis anteriores, recurrirá a la cognición para protegerse en situaciones sociales críticas. Esto haría de la cognición un *escudo protector* que presumiblemente se apaga más rápido de lo que uno se puede dar cuenta y ya no termina en la flexible zona de control del escritor, sino que se asemeja a un automatismo. Los escudos protectores tienen funcionar y uno no piensa en si los utiliza o no. Para la terapia, esto significa que uno puede entonces iniciar el cambio en el escritor *en el momento en que se levanta el escudo* y, a partir de entonces, cuando estalla la crisis. Una nueva acción se aprende mejor cuando es necesaria. De lo contrario más bien en seco. Para el cambio estructural, se necesita la nueva experiencia de cambiar en la crisis, lo que contradice las cogniciones que se ejecutan automáticamente. Sin duda se necesitan repetidas experiencias nuevas para que se registren siquiera en el sistema del cliente potencial. De lo contrario, hay una falta de confianza para comprometerse con la propia, ciertamente vulnerable emocionalidad e iniciar la recuperación y la curación interior. Así pues al nivel externo, las cogniciones son un recurso que sin duda puede utilizarse para la organización, la formación, etc. y que ya parece estar bien entrenado – puesto que está activo como escudo protector – todavía le queda un camino muy largo por recorrer en lo que respecta a la emocionalidad y los componentes de acción motivacional asociados. *Emocionalmente*, este cliente potencial necesitaría mucho apoyo y un salto de fe para desarrollar la confianza y comprometerse con el cambio.

Lo dejaremos en este punto con el análisis de la primera fase del análisis de la secuencia. La estructura del caso está lejos de repetirse, lo que normalmente es señal de la finalización del proceso de generación de una hipótesis preliminar de estructura del caso. En este punto observamos tres componentes significativos con los que formamos la hipótesis preliminar de la estructura del caso:

- Presencia de recursos cognitivos que contradicen la
- emocionalidad débilmente apoyada en el contexto de
- la evitación de la concreción y múltiples dispositivos estilísticos de ocultación.

Al mismo tiempo, en relación con la historia biográfica del escritor, sospechamos

- experiencias de ser ignorado, no suficientemente apoyado emocionalmente y de ser excluido o segregado,
- y posibles traumas que deben especificarse con más detalle.

Estas experiencias conformaron de manera significativa la figura básica de la estructura del caso esbozada anteriormente. Nos mantenemos deliberadamente alejados de las atribuciones mono-causales, ya que rara vez son aplicables. Hay demasiadas y ya prenatales influencias en la vida como para que tenga sentido, desde nuestro punto de vista, abalanzarse sobre un único o muy pocos acontecimientos estrechamente definidos sólo para explicar un problema tan amplio como la adicción. Nosotros partimos de acontecimientos multi-causales complejos y a largo plazo. Este punto de vista puede entenderse bien bajo el aspecto de la teoría de la oscilación: Algunas oscilaciones se anulan entre sí, otras se suman hasta amontonarse, escalar y desarrollar fuerzas muy grandes, determinando o incluso destruyendo todo lo demás, y otras ondas no se estorban entre sí y, por tanto, prácticamente no se influyen entre sí.

Después de esta hipótesis preliminar de la estructura del caso, se requiere el paso de la falsificación.

11.13.2.3 La falsación de la hipótesis preliminar de la estructura de casos

Pasamos ahora a un intento de falsación de la hipótesis preliminar de la estructura de caso. Al final de la última sección, se enumeraron tres elementos más un posible factor de influencia significativo de la experiencia biográfica del cliente potencial. La tarea de falsificación consiste ahora en seleccionar una frase o parte de una frase que parezca adecuada para demostrar empíricamente lo contrario de nuestras afirmaciones basándonos en el texto. Por lo tanto, no seguimos el curso de la frase que aún no ha sido

analizada hasta el final, porque entonces existe el peligro de que continúe como suponemos, pero más tarde resulte muy diferente. Nos decidimos por el último conjunto. Esto debería dejar una impresión como conclusión comparable al efecto de primacía como efecto de recencia y resumir el objetivo de la carta de forma repetida y condensada. Hasta aquí la teoría: veamos la realidad.

En primer lugar, consideremos cómo podemos reconocer una falsación de la forma más clara posible. No existen, como en el contexto de los valores numéricos, umbrales críticos que separen lo significativo de lo insignificante. Incluso éstos están sujetos a error y no son necesariamente inequívocos, como muestra el ejemplo del ROPE (véase el capítulo 6.8.4.2), o el hecho de que la diferencia entre "significativo" y "no significativo" no sea en sí misma estadísticamente significativa (Gelman y Stern, 2006). La estadística es un buen ejemplo para ilustrar el problema de lo que significa encontrar información fundamentada en el material que separe lo significativo de lo insignificante y nos permita así sacar conclusiones sobre el proceso global, la hipótesis de la estructura del caso, actualmente aún preliminar. También debemos ser conscientes de que sólo se trata de un modelo si, como en el razonamiento binario, clasificamos o interpretamos claramente una hipótesis cualitativa como confirmada o no confirmada, es decir, falsada. Aunque, por supuesto, esto puede practicarse de forma binaria con *confirmado frente a no confirmado*, nosotros suponemos que la percepción humana funciona más como una distribución posterior bayesiana en la que se superan umbrales para que una diferencia cuantitativa se convierta en una diferencia cualitativa, pero al mismo tiempo existe una zona de incertidumbre en la que conviven lo significativo y lo insignificante. Además, partimos de la base de que los umbrales críticos no son necesariamente unívocos siempre y en toda circunstancia, sino que cambian de forma dinámica y en función de la situación, y posiblemente también con el paso del tiempo.

No obstante, nos comprometemos – reconocemos una falsación de nuestra hipótesis preliminar de estructura de casos al statu quo por las siguientes características, entre otras:

- Minimizar el componente cognitivo o ponerlo directamente al servicio de la emoción, la motivación y la acción. Hasta ahora, el curso se ha caracterizado por el predominio de actividades cognitivas como la imaginación o la lectura de conceptos. Por tanto, la falsificación requiere una reducción masiva de este tipo de actividades en el texto.
- Expresión de emocionalidad estable, es decir, expresar emociones como la alegría o el miedo en el contexto de una conclusión estándar esperada, por ejemplo, formalmente correcta: "Me alegraría mucho tener noticias suyas/invitación a una entrevista personal". Abordar directamente el fracaso y las emociones asociadas a él también sería falsificable. También sería falsificador si se discutieran pasos de acción concretos sobre una posible permanencia en terapia o sobre la propia recuperación. También sería una falsificación si el escritor relatara experiencias concretas anteriores de naturaleza emocional o con una connotación emocional, por ejemplo (versión larga) "Estoy muy desesperado. Quiero desesperadamente un lugar de terapia contigo pronto, porque sé que pronto puedo matarme con mi estilo de vida actual, y eso me da mucho miedo".
- Encuentro y concreción, es decir, no se habla de la institución de forma inespecífica, sino que se concreta o se hace sugerencias, por ejemplo, sobre una entrevista. Por cierto, no se trata de una expresión de pseudo-autonomía, ya que el principio de la capacidad de fracaso implica que las propuestas pueden ser rechazadas o modificadas por la institución. Al contrario, darle forma, intentarlo, es una expresión de autonomía (residual) efectiva en la acción. Se trataría entonces de una expresión de encuentro y comunicación entre iguales, teniendo en cuenta que los participantes no son tan iguales y que el desequilibrio de poder ciertamente no está del lado de quien escribe.
- Alejándose de la evasión del yo y del subjuntivo, es decir, el yo se muestra como actuante en la realidad y se responsabiliza de ella. Incluso una exigencia destemplada como "espero tu respuesta pronto para poder aceptar inmediatamente la plaza contigo" representaría una clara falsificación de los supuestos actuales.

Por cierto, una confirmación de nuestra hipótesis propuesta sobre la estructura del caso esperaría idealmente que la carta de solicitud terminara exactamente como empezó. Sin embargo, una pequeña variación respecto al caso ideal no conduce directamente a la falsación. Aquí es donde surge el problema de los umbrales para separar significación y despreciabilidad mencionado anteriormente, que desde nuestro punto de vista encontramos tanto en lo cualitativo como en lo cuantitativo. Falsificación significa que aparecen elementos que ya no son compatibles con la hipótesis preliminar de la estructura del caso. Las ligeras variaciones en los detalles no invalidan inmediatamente las hipótesis básicas, pero pueden significar tener que re-elaborar los detalles con respecto a determinadas cuestiones. No se puede abarcar todo y todos los ámbitos de la vida con una estructura de casos, pero sí los ámbitos temáticamente relevantes. Nuestra

hipótesis preliminar de la estructura de caso es breve y sencilla en su versión corta, pero por esta misma razón no es lo suficientemente detallada como para predecir todos los casos y todas las pequeñas variaciones. Es: *emocionalidad débilmente apoyada con fuerte expresión de cogniciones y evitación de la concreción y la experiencia directa*.

La pasividad mencionada y la ocultación de uno mismo como agente es bastante típica de los drogodependientes, cuya principal tarea de aprendizaje suele ser responsabilizarse de los propios actos y desprenderse del egocentrismo radical. Pasar constantemente a una postura de pasividad es una expresión de un egocentrismo mal entendido muy alto, porque se suspende la actuación según las necesidades situacionales. Al mismo tiempo, el error de atribución causal tiene lugar todo el tiempo: los fracasos se atribuyen a la situación y a fuerzas externas en lugar de relacionarse con uno mismo en el sentido positivo de de asumir la responsabilidad e independientemente del resultado de los acontecimientos. Sin experimentarse a uno mismo como agente con resultados positivos, naturalmente hay problemas a nivel de motivación para cambiar. No en vano el conductismo dice "aprender por el éxito" y no por el fracaso o la imaginación, sino por el comportamiento. En una actitud puramente pasiva uno mismo queda totalmente excluido y, por tanto, no puede ser responsable, a la inversa, no se experimenta a sí mismo como agente y, por tanto, no puede acumular éxitos y, desde luego, no puede fracasar para hacerlo mejor en el futuro. Como sabemos por la teoría de la ciencia, el desarrollo del conocimiento requiere el fracaso como única forma de aprender algo nuevo. Si ocurre algo que ya se espera y no se resiste, no se aprende nada. Esta situación es extremadamente difícil e impide el desarrollo. Por cierto, tal hipótesis estructural casuística no significa, por supuesto, que esta estructura no pueda disolverse y cambiar fundamentalmente en el curso de la terapia y diversas experiencias nuevas.

Los elementos del caso ideal para apoyar la presente hipótesis estructural casuística preliminar son, pues, para esta última frase

- orientación cognitiva principal de la frase y sus elementos
- falta de emocionalidad clara y posicional
- evitación de abordar acciones concretas reales, planes y emociones asociadas
- ocultamiento de sí mismo, del "yo", en la pasiva del modo subjuntivo
- seguir dirigiéndose a la institución en primera persona.
-

Para la falsificación, miramos toda la última frase en una y no de unidad más pequeña a unidad más pequeña o de signo de puntuación a signo de puntuación:

```
Espero que me concedais esta oportunidad a pesar de
mi edad relativamente avanzada.
[Ich hoffe, dass Ihr mir diese Chance trotz
meinem relativ hohen Alter gewährt.]
```

La frase comienza de forma comparable a la primera frase. "Tengo [...]" se une a "Espero [...]", "vuestro" se une a "concedais". El término "esperanza" implica la creencia o expectativa de un resultado positivo de un acontecimiento en el futuro. La esperanza no garantiza que este acontecimiento vaya a tener un resultado positivo, por lo que contiene una gran incertidumbre que no necesariamente puede reducirse con las propias acciones. Por lo tanto, el resultado del acontecimiento escapa al propio control y esfuerzo. Podríamos preguntarnos qué hace el cliente potencial concretamente y cuánto se esfuerza cuando sabe que poco puede hacer por el resultado. Esto no significa que él no pueda hacer nada en absoluto. Tradicionalmente, ser "de buena esperanza" significa estar embarazada y esperar un hijo, y ya se ha hecho mucho de antemano. El dicho procede de una época en la que la medicina estaba menos presente que hoy y la religión seguía siendo muy influyente. Y las religiones, como sabemos, almacenan lo esencial de la existencia en lo numinoso, que no puede ser controlado en absoluto ni directamente por el ser humano.

Podemos interpretar este comienzo de frase como una expresión máxima de emoción oculta. Esto significa que el comienzo de la frase es bastante congruente con lo que se ha dicho hasta ahora. Falta una declaración emocional clara. Incluso un énfasis como "muy", "realmente", etc. en relación con la esperanza no lo encontramos. Por tanto, la hipótesis de la emocionalidad oculta no se falsifica, sino que incluso se refuerza.

La segunda parte de la frase que comienza con "que [...]" suena construida y, en consecuencia, de nuevo muy cognitiva. Aunque es correcto afirmar que se conceden [gewähren] oportunidades, en alemán normal y corriente se diría "recibir", "dar" o "permitir" y no "conceder". El verbo "gewähren" pertenece al sustantivo "Gewähr" y significa "garantia", en parte "cobrar" y tiene el aspecto de "preservar" en el contexto de las ideas morales y religiosas. El origen denominativo de "gewähren" se encuentra en el significado del sustantivo "Gewähr/garantía", que aún hoy se encuentra en la garantía legalmente regulada para la compra de productos. La garantía estipulada legalmente contrasta con la declaración voluntaria de garantía por parte de un fabricante. La garantía es un aseguramiento jurídicamente vinculante de la calidad de un artículo. En el caso de los números de lotería o los horarios de trenes, la garantía se excluye explícitamente para proteger contra reclamaciones innecesarias. Sin embargo, esto aquí no es un asunto legal y la posibilidad de un compromiso legal no sólo suena complicado, sino más bien como un lío legal o al menos mucho esfuerzo. Al final, esta redacción lo complica mucho más de lo necesario. Y todo esto apoya las suposiciones anteriores sobre el papel de las cogniciones en el cliente potencial. Precisamente nuestra suposición de que las cogniciones afloran automáticamente como escudo protector en el escritor en una crisis no queda invalidada por esta parte del texto, sino reforzada.

Al fin y al cabo, la oportunidad mencionada podría ser la última y el fracaso podría conducir como acontecimiento a la crisis máxima, a una sobredosis y, por tanto, a un suicidio al menos indirecto. Esta crisis no es puesta en evidencia directamente por la cognición, sino que se empaqueta y expresa de forma codificada. Así que podemos preguntarnos seriamente, ¿hablaríamos de "concesión" ante "la" posibilidad de una plaza en terapia? Este término crea una gran distancia con respecto a la experiencia y se refiere a regulaciones legales y leyes sobre las que el individuo tiene poca o ninguna influencia ad hoc en términos de configuración. Pero aquí no hay ni garantía de plaza terapéutica ni garantía de éxito terapéutico. La primera depende de muchos factores y la segunda especialmente de la evolución y los cambios por parte del propio cliente. El autor suena como si quisiera asegurarse la plaza de terapia en cuanto a regulación legal, pero al mismo tiempo sigue entendiéndola como una oportunidad y no la exige insolentemente en el sentido de una garantía. Intuitivamente, tal expresión suena desesperada en esta situación. Pero las emociones no se abordan ni se expresan directamente. El cliente se siente muy mal en la última frase, pero no puede expresarlo.

El artículo directo "diese/este" antes de "Chance/azar" puede interpretarse como un mayor énfasis de la emocionalidad vulnerable oculta. El escritor señala así que sabe muy bien que no tiene demasiadas opciones de acción en este momento y que la institución aquí presente puede darle o permitirle no sólo "una" oportunidad, sino "una de las pocas" oportunidades en absoluto, o incluso "la última". La tesis de la emotividad, que sólo funciona de forma indirecta o apenas es visible porque está débilmente apoyada, emerge claramente y es, en rigor, incluso una discreta desviación de lo completamente inespecífico del principio de la carta. Podría haber escrito podría haber escrito "*una* oportunidad"; pero eligió "*esta*". Este es un buen punto de partida, ya que podemos tomar este punto como el mínimo denominador de la auto-reflexividad. El cliente potencial parece ser consciente de que no tiene muchas más oportunidades. Esto se ve reforzado por la afirmación relativizadora adjunta de "alteridad relativamente alta". Podría haber escrito de otra forma: "Me doy cuenta de que soy viejo, pero le ruego que me dé esta oportunidad". Por tanto, la construcción oracional elegida se asemeja de nuevo a a un elemento abstracto-cognitivo no concreto que ya hemos tratado suficientemente.

No vemos una clara falsificación, sino más bien un apoyo más a la suposición de que el intelecto del cliente potencial interviene y lo hace más complicado y difícil de lo que realmente es. En otras palabras, el intelecto necesitaría una base para actuar por la totalidad, por todo el ser humano. Pero esta base, las emociones, no son estables y no pueden expresarse. Así que el intelecto actúa sin una base emocional, y esto acaba en formulaciones que ni parecen sencillas ni realistas y, sobre todo, parecen desvinculadas del resto del sistema.

El pasaje "a pesar de mi edad relativamente avanzada" tiene varios aspectos. En primer lugar, es una información importante por su contenido. Para ser precisos, desgraciadamente no sabemos si en aquel momento la institución conocía la edad exacta del cliente, pero la suponemos. El cliente es efectivamente "viejo" para un adicto en el momento de la solicitud, a saber, 38 años. Para la duración normal de la vida de un ser humano, todavía no es muy viejo. En este sentido, la relativización "edad ... avanzada" es una descripción adecuada de esta situación. Hay que relacionar esto con la información de que a menudo se

seleccionan personas presumiblemente más jóvenes que las "cronificadas" cuando las plazas terapéuticas son limitadas. Éticamente, no hay ninguna razón para ello, pero desde luego no es raro. Esto sugiere que el escritor tiene experiencia previa de no ser aceptado como cliente debido a su edad. El contraste que comienza con "oportunidad" y luego "a pesar de" y "edad" se inserta en la frase antes de que el escritor pase a "conceder". Esta construcción confiere a la frase una gran urgencia. Pero al mismo tiempo, el autor vuelve a debilitarla antes de que pueda cobrar sentido. Podría escribir "Soy relativamente viejo y por eso necesito esta oportunidad, porque muchas instituciones ni siquiera me miran". Así, podría utilizar directamente un punto débil, su edad, como argumento para recomendarse para la plaza de terapia: *Utilizar las vulnerabilidades como recursos sería el lema entonces* (véase Gürtler, Studer & Scholz, 2012). Al fin y al cabo, lleva consigo cierta resistencia y robustez a pesar de seguir vivo a pesar de su avanzada edad y su gran consumo de drogas. Podría mencionar esto añadiendo que su robustez también puede acabar muy rápidamente. En lugar de ello, vuelve a transformar todos estos aspectos y hace de ello una frase indirecta, que carece de un encuentro con la otra persona. Podría haber apelado a la conciencia moral de la institución o de la dirección o de los terapeutas, como con un "¡Alguien tan viejo como yo podría morir pronto, tenéis que ayudarme!" En su lugar, repite la figura de la cognición que ya conocemos, encubriendo la emocionalidad oculta y vulnerable, y la vuelve a conectar con la falta de concreción.

La percepción de la relevancia de esta oportunidad en el contexto de la propia edad y la experiencia previa ciertamente existente de solicitar terapias y solicitar plazas de terapia habla en favor de una alta motivación para la terapia. Esto no contradice la tesis de una emocionalidad débilmente apoyada y una orientación cognitiva excesivamente alta. Todas estas experiencias previas, la elevada edad y la motivación terapéutica resultante pueden utilizarse en la terapia. Sin embargo, el escritor no parece ser consciente de ello, ya que no utiliza nada de ello directamente para recomendarse a sí mismo.

Basándonos en esta frase, no vemos invalidada nuestra hipótesis preliminar de la estructura del caso, sino más bien confirmada. Sin embargo, aún vemos trabajo por hacer para elaborar la relación entre emocionalidad y cognición de forma más precisa. Del mismo modo, todavía requiere cierto esfuerzo determinar la forma exacta en que el escritor, en combinación con sus propias emociones, se oculta y en un papel pasivo, del texto. Y esto es sólo el principio. Faltan los análisis de la biografía familiar a través del genograma, los posteriores registros de terapia y el curso de la vida post-catamnésica. Con la ayuda de las entrevistas post-catamnésicas existentes, se podría examinar si la estructura del caso aquí establecida sigue siendo válida años después de la terapia. Esto no sería una falsificación, sino una prueba de la eficacia de la terapia contra la adicción.

Tarea 11.1: Ampliación y falsación de la hipótesis de la estructura de caso

La falsación más profunda de la presente hipótesis de estructura de casos o incluso la ampliación de la propia hipótesis de estructura de casos sería una tarea para lectores comprometidos y con tiempo suficiente. Hay material suficiente para trabajar con detenimiento y ampliar más la hipótesis de la estructura de caso con el fin de someterla a múltiples intentos de falsación.

También sería posible trabajar con material de los expedientes terapéuticos, que no podemos revelar por razones de protección de datos. Los lectores pueden tomar el material sobre el cliente impreso en Studer (1998) (por ejemplo, el genograma) más la información sobre el curso post-catamnésico (Gürtler, Studer & Scholz, 2012). El material es siempre completamente anónimo y de acceso público. Con su ayuda, se puede intentar invalidar la hipótesis de la estructura del caso en cuestión o primero desarrollarla. Por supuesto, ambos pasos no deben llevarse a cabo sobre el mismo material. De lo contrario, se abandona el camino de la ciencia.

En el caso de la falsificación, se procede como se ha descrito anteriormente, en el sentido de que los pasajes críticos del material son material con el fin de invalidar la hipótesis de la estructura del caso y demostrar lo contrario. En primer lugar, se debe determinar por escrito y de forma muy concreta cuándo se produce la falsificación y, a la inversa, qué habla en favor de una confirmación de las hipótesis existentes. Se trabaja según principios analíticos secuenciales, según los cuales, en el caso de los intentos de falsificación se permite explícitamente seleccionar los pasajes de texto relevantes. Por lo tanto, se toman los pasajes críticos o poco claros o dudosos. En caso de duda y hipótesis de la estructura del caso bien elaborada, estos pasajes críticos apoyarán la hipótesis en cuestión en lugar de invalidarla.

11.13.2.4 Reflexión sobre la hipótesis (preliminar) de la estructura de caso

Sólo hemos interpretado algunas líneas y de ningún modo hemos analizado la estructura del caso hasta el punto de repetirla en su totalidad. Esto se debe únicamente a que queríamos limitar el esfuerzo y, sin embargo, ahora tenemos un texto muy largo. Las argumentaciones, que hemos formulado deliberadamente al estilo del pensamiento en voz alta, muestran la cantidad de material que puede producir incluso un pequeño pasaje del texto. El análisis de secuencias nos ofrece acceso a una especie de microscopio con el que se pueden identificar estructuras muy finas con gran energía, de modo que con él se pueden reconstruir con precisión todos los demás aspectos del caso, como con un holograma. Otra forma de ver el análisis de secuencias es la del cocodrilo que abre la boca, permite todas las interpretaciones y luego pasa al control de la realidad para el siguiente pasaje del texto. Entonces la boca se cierra y todo lo que no tiene sustancia sale o desaparece en la garganta del cocodrilo. Entonces el cocodrilo vuelve a abrir la boca y el procedimiento se reproduce hasta que la estructura del caso se repite una vez y no sale nada nuevo y lo que queda realmente tiene sustancia. Esta sustancia resiste por sí misma nuevos intentos obstinados de ser aplastada o tragada cuando se cierra la boca. Después de eso, el cocodrilo ha tenido su día por ahora y puede dedicarse a otros temas, porque entonces una hipótesis preliminar de estructura de caso se habrá logrado no falsar y se podrá utilizar como estructura de caso.

Podemos formular una hipótesis de estructura de caso sobre el caso anterior, pero no la hemos falsado de forma suficientemente específica y amplia, sino sólo sobre un único pasaje del texto. En la práctica de la investigación, esto no sería suficiente. También se necesitarían urgentemente otras fuentes de información para evitar ignorar accidentalmente más contextos relevantes del caso que los que proporciona una sola carta de solicitud. No hay que subestimar el tiempo que requiere un análisis de este tipo. Pero como resultado, se ha elaborado realmente algo sustancial que perdurará.

Llegados a este punto, lo dejamos en la hipótesis hasta ahora no falsada de la estructura del caso, que se lee de forma muy abreviada: *emocionalidad débilmente apoyada con pronunciadas capacidades cognitivas y una fuerte tendencia a evitar el encuentro del yo como agente*. La preliminar hipótesis tiene el potencial de una estructura de caso si se completan hasta el final los pasos empíricos necesarios enumerados.

11.14 Discusión: Hermenéutica Objetiva

El análisis de secuencia es un procedimiento complejo con muchas reglas relativamente estrictas que hay que seguir. El marco teórico subyacente de la Hermenéutica Objetiva ofrece, por un lado, amplia orientación teórica para la práctica de la investigación en ciencias sociales más la justificación práctica de la investigación, pero requiere un largo periodo de formación y una supervisión inicial por parte por investigadores formados. Las traducciones a otros idiomas son prácticamente inexistentes. Por último, hay la cuestión si toda la teoría de la Hermenéutica Objetiva es siempre, o incluso necesaria en absoluto para poder aplicar seriamente el análisis secuencial en contextos prácticos y científicos.

Sin embargo, desde nuestro punto de vista, esto no tiene por qué ser así. En una gran variedad de talleres y cursos de formación, así como en la práctica, que con un mínimo información y adecuado control metodológico junto con los "profanos" se pueden obtener resultados muy razonables y bien fundados. Claro, lo que se necesita es la supervisión inicial, o al menos el trabajo en grupo (supervisión entre iguales). Es importante deshacerse de la timidez del proceso y, a través de una práctica cada vez mayor, mejorar constantemente la calidad del propio trabajo y aprender cumplir las numerosas normas por su cuenta.

El uso del ordenador (Gürtler & Huber, 2016) sin duda facilita, especialmente a los principiantes, el trabajo con el método, ya que los errores típicos (como las violaciones de la secuencialidad) se evitan o son evitados o minimizados por el software. Otros errores, entre los que se incluyen errores de naturaleza interpretativa, así como el enfoque en la literalidad (véase cap. 11.9.6) no pueden ser interceptados por el software. Lo único que ayuda en este caso es cuestionar críticamente el propio trabajo, dejarlo todo por un día y mirarlo con otros ojos, sin olvidar la supervisión por pares, que puede evitar a tiempo muchos errores en el proceso de análisis.

En el lado positivo, sin embargo, se encuentra un método increíblemente potente de análisis de datos cualitativos que se puede utilizar en prácticamente todos los contextos y que consideramos extremadamente eficaz a pesar del esfuerzo obviamente elevado que implica. No querríamos prescindir de él. Al contrario, hemos empezado a incorporar las reglas del análisis de secuencia al paradigma de codificación como parte del paradigma de codificación, con el fin de asentar los códigos sobre una base bien fundamentada en una fase temprana, por ejemplo, si posteriormente se van a realizar análisis estadísticos o lógicos. De estos últimos que nos ocupa ahora.

Parte IV

Métodos lógicos

Capítulo 12
Minimización Booleana o Análisis de Implicantes

> »It appeared to me that, although Logic might be viewed with reference to the idea of quantity, it had also another and a deeper system of relations.«
>
> *The Mathematical Analysis of Logic, 1847, Preface*
> George Boole, 1815–1864

12.1 Propedéutica

En realidad, esto no difiere de los métodos cuantitativos, que intentan ofrecer afirmaciones y conclusiones lo más precisas posibles, tanto a nivel grupal como individual. Todo esto suena un poco a la búsqueda del perpetuum mobile.

Así que queremos comparar sin perder demasiada información y detalles. Por eso hacemos estadísticas, entre otras cosas. La cuestión es: ¿podemos conseguir ambas cosas al mismo tiempo? ¿O sea, así que va un *ambos/y* en vez de un *o/y*? ¿Podemos combinar estas dos perspectivas obviamente opuestas de forma complementaria para obtener un resultado global para nuestro proyecto de investigación? Se trata de añadir un análisis comparativo precisamente de estos casos individuales a diversos análisis de casos individuales y sus posibles peculiaridades subyacentes. Con ello, perseguimos el objetivo de reducir los casos individuales al mínimo posible y obtener los resultados de los mismos. El siguiente capítulo ofrece una forma de realizar esta intención, lejos de codificaciones, interpretaciones y extractos numéricos: trabajando únicamente con *conexiones lógicas*. El trabajo con operaciones lógicas fue creado por George Boole (1815-1864) a mediados del siglo XIX y constituye la base de todo ordenador o smartphone actual, que se basa en la lógica binaria [0, 1]. Y de eso se trata exactamente: de hacer comparaciones que sean lógicamente `VERDADERAS` o lógicamente `falsas`. Lógicamente `VERDADERO` o lógicamente `falso` no significa nada a nivel de contenido. No se trata de una evaluación de códigos o interpretaciones, sino de un proceso abstracto de comparación según las reglas de la *lógica formal*, que se ocupa de las conexiones entre enunciados y las conclusiones que se pueden derivar de ellos. Por ejemplo, en el caso de las comparaciones lógicas AND (conjunción, código R en `ptIV_qual_Boole_basics.r`) se aplica lo siguiente

```
> TRUE & TRUE
[1] TRUE
> TRUE & FALSE
[1] FALSE
> FALSE & TRUE
[1] FALSE
> FALSE & FALSE
[1] FALSE
```

y en el caso de las comparaciones lógicas OR (disyunción, código R)

```
> TRUE | TRUE
[1] TRUE
> TRUE | FALSE
[1] TRUE
> FALSE | TRUE
[1] TRUE
> FALSE | FALSE
[1] FALSE
```

Tabla 12.1: *Tablas de valores de verdad (operaciones lógicas de un dígito de "A" y "A")*

no.	Entradas {A, A}		Salida
	A	A	
1	0	0	0
2	0	1	1
3	1	0	0
4	1	1	1

Mientras que el AND lógico consiste en que toda la expresión es lógicamente VERDADERA si todos los componentes son verdaderos, lo contrario es cierto para el OR lógico -- la expresión es lógicamente VERDADERA si uno de los componentes ya es verdadero:

```
> FALSE | FALSE | FALSE | TRUE | FALSE
[1] TRUE
```

El orden no importa, a menos que se introduzcan paréntesis que correspondan a la habitual priorización algebraica:

```
> (FALSE | FALSE | FALSE | TRUE | FALSE) & (TRUE & TRUE)
[1] TRUE
> (FALSE | FALSE | FALSE | TRUE | FALSE) & (TRUE & FALSE)
[1] FALSE
```

De forma equivalente, existen otras operaciones lógicas que permiten construcciones bastante complejas. Las operaciones lógicas más comunes de uno y dos dígitos se enumeran en las tablas 12.1 y 12.2. Generalmente, por convención, un "0" significa lógicamente `falso`, un "1" lógicamente VERDADERO. Trasladado a nuestro objetivo de *generalización con la mínima reducción de información*, esto significa que las comparaciones lógicas no tienen lugar simplemente en un nivel muy abstracto, sino siempre en el nivel concreto que nosotros mismos denotamos. De este modo, podemos generar comparaciones de categorías y encontrar así un nivel de abstracción adecuado. De este modo se tiene en cuenta la tipicidad subyacente en un nivel de abstracción calibrable para que las comparaciones puedan ser positivas y no queden oscurecidas por la unicidad, y viceversa. La calibración surge de la elección del nivel de abstracción de los códigos (condiciones) y está relacionada con el contenido. Su objetivo es producir condiciones comparables entre casos, estudios, etc. Si no se consigue, los casos no pueden compararse entre sí o sólo con grandes compromisos. Un ejemplo es el conocido estudio de Krook (2010), que examina qué factores determinan la representación de las mujeres en los parlamentos. Se comparan países occidentales con países del África subsahariana. Desgraciadamente, para estos dos bloques se utilizan condiciones diferentes – justificables en cuanto al contenido –, que se asocian al criterio. A pesar de la justificabilidad de fondo de la selección de las condiciones, esto conduce al hecho de que una comparación directa "Estados occidentales frente a los Estados del África subsahariana" no puede realizarse de este modo. La *selección de la muestra* y la *creación*

de categorías relevantes para la comparación requiere siempre un esfuerzo cuidadoso para evitar estas situaciones.

Tabla 12.2:*Tablas de valores de verdad (operaciones lógicas de dos dígitos de "A" y "B")*

no.	Vinculación de A y B		Entradas {A, B}			
	A		0	0	1	1
	B		0	1	0	1
	Definición de la función		Salidas {A, B}			
1	Contradicción		0	0	0	0
2	Conjunción (AND)		0	0	0	1
3	Inhibición (AND y negación inicial)		0	0	1	0
4	Identidad de A		0	0	1	1
5	Inhibición (AND y negación inicial)		0	1	0	0
6	Identidad de B		0	1	0	1
7	Antivalencia (eXclusivo-OR = XOR)		0	1	1	0
8	Disjunción (OR-Vínculo = OR)		0	1	1	1
9	Función de Peirce- (No-OR = NOR)		1	0	0	0
10	Equivalencia (No-eXklusivo-OR = NXOR)		1	0	0	1
11	Negación de B		1	0	1	0
12	Implicación de B (OR y negación inicial)		1	0	1	1
13	Negación de A		1	1	0	0
14	Implicación de A (OR y negación inicial)		1	1	0	1
15	Función de Sheffer (No-AND = NAND)	no ambos	1	1	1	0
16	Tautología		1	1	1	1

Para nuestro tema de *búsqueda de configuraciones condicionales mínimas* bastarán los enlaces binarios de dos dígitos que se muestran en la Tabla 12.2. La tabla debe leerse de forma que los respectivos enlaces de dos dígitos de A y B conduzcan a una salida específica. Por ejemplo, para la contradicción (ejecución 1), los enlaces de A = 0 = VERDADERO y B = 0 = falso conducen al resultado de la contradicción = 0 = falso. Si se observa la tabla en su conjunto, se observa que abarca sistemáticamente todas las posibilidades combinatorias y cada posibilidad de enlace recibe un nombre. Con un enlace de dos variables, cada una de las cuales puede tener dos expresiones VERDADERO [= 1] y falso [= 0], surgen 2^4 = 16 combinaciones posibles, que corresponden exactamente a las 16 funciones de la tabla 12.2.

Recordatorio 12.1: Paradojas de la metodología cualitativa

El dilema al trabajar con métodos cualitativos consiste a menudo en querer preservar la unicidad y al mismo tiempo hacer posibles las generalizaciones, es decir, la reducción de la información sin una reducción "real".

La realización de las comparaciones tiene lugar mediante software, por ejemplo con el software de código abierto AQUAD 7 (Huber, 2019) o R (R Development Core Team, 2019d). Por supuesto, existen enlaces de orden superior, pero estos pueden expresarse en su mayoría mediante una combinación de enlaces de dos dígitos, así como la posibilidad de no aplicar una lógica binaria [1, 0], sino una que pueda asumir más de

dos estados diferentes o permita gradaciones. Ejemplos de ello son los análisis de valor añadido de lógica difusa (Fuzzy Logic, s. cap. 12.7), que dividen el espacio binario entre lógicamente VERDADERO y lógicamente FALSO en múltiples categorías (= manifestaciones de una propiedad). Suponen una generalización del álgebra booleana binaria a más de un estado binario por condición, propiedad, etc. Además, la pertenencia definitiva cambia a pertenencia parcial con respecto a un estado; es decir, los conjuntos son difusos. Los conjuntos ya no se denotan por sus objetos, que les pertenecen o no, sino por el grado de pertenencia a un conjunto. Básicamente, esto significa que la pertenencia a un estado puede expresarse como una probabilidad. Una función de pertenencia asociada modela numéricamente el conjunto. La diferencia con la estadística es que se utilizan operaciones con conjuntos lógicos y no cálculos estadísticos. Las operaciones con conjuntos se basan en el grado de pertenencia a un estado.

Para los *análisis comparativos cualitativos*, la lógica binaria suele ser suficiente, pero, como siempre, la elección del procedimiento de análisis debe estar en consonancia con la pregunta de investigación y los datos de que se trate o sus especificidades. Los propios vínculos lógicos pueden representarse a menudo mediante diagramas de Venn a partir de la teoría de conjuntos, lo que facilita la comprensión de las operaciones lógicas (véase un ejemplo en la Fig. 12.1). A partir de entidades de seis o siete dígitos, el diagrama de Venn se vuelve tan confuso que ya no puede representarse razonablemente. E incluso antes de eso es bastante confuso, aunque sea preciso.

Ésa es la idea básica de los métodos comparativos. Ahora recurrimos a las ideas básicas de la *minimización booleana* para identificar conjuntos mínimos de condiciones para los cuales un fenómeno se convierte en lógicamente VERDADERO o lógicamente FALSO.

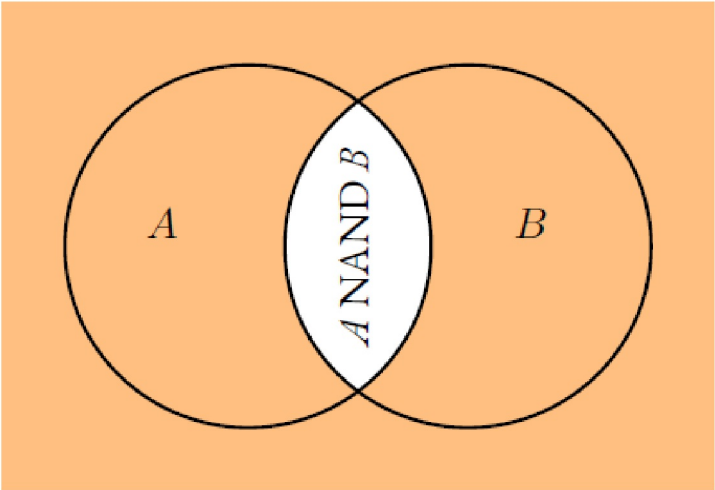


Figura 12.1 *Diagrama de Venn (vinculación NAND)*

12.2 Ideas básicas de la minimización booleana

Las comparaciones constantes de interpretaciones dentro de un conjunto de datos (texto, vídeo, grabación de audio, imagen) y entre distintos conjuntos de datos constituyen el núcleo de los procedimientos de análisis cualitativo (Shelly y Sibert, 1992). Ya en el caso de la reducción de datos dentro de un archivo, sin comparaciones de las categorías o los segmentos de datos tanto entre sí como entre los otros archivos, no se puede lograr una codificación fiable. El resultado sería un sistema de codificación tan único que apenas podríamos descubrir puntos en común entre categorías o incluso entre casos. El beneficio sobre el texto original sería nulo.

Ahora bien, en la mayoría de las investigaciones, uno quiere captar la singularidad de cada conjunto de datos o las opiniones, competencias, visiones subjetivas del mundo, etc. expresadas en él no solamente con un sistema de categorías válido para todos los expedientes. En algún momento del proceso de investigación, uno quiere elaborar configuraciones generales para todos los expedientes. Por tanto, en la fascinación por la originalidad y la singularidad, hay que lidiar con el proverbial que a menudo los árboles no dejan ver el bosque. Sobre todo diferencias y su especificación, no hay que perder de vista el hilo común. Para ello, hay que reconocer y comparar las conexiones sistemáticas entre los archivos. Sin embargo, el problema de comparar contextos con fenómenos sociales es que a menudo se encuentran una multitud de condiciones en combinaciones diversas, a veces incluso contradictorias. Esto se aplica a todos los niveles de nuestros análisis cualitativos hasta ahora dentro del paradigma de codificación (véase el capítulo 9). Al fin y al cabo, la tarea del análisis es precisamente aclarar estas diferencias.

Recordatorio 12.2: Lenguaje, codificación e hipótesis

Lenguaje original – > códigos conceptuales a cualquier nivel de abstracción (contenido "¿qué?" y estructura "¿cómo?") – > Vinculación de hipótesis

En muchos estudios, uno se enfrenta a la tarea adicional de comparar sus propios resultados con otros estudios, es decir, básicamente realizar un *metaanálisis cualitativo*. Aunque otros investigadores hayan trabajado sobre cuestiones comparables sólo habrán encontrado respuestas parcialmente coherentes. En todas las ciencias que se ocupan de la complejidad de los sistemas naturales, la búsqueda de relaciones causales para un determinado fenómeno encuentra una gran variedad de constelaciones a través de diversos estudios. Tomemos una pregunta a menudo agriamente debatida por los adversarios en la vida cotidiana como ejemplo: "¿Existe una relación entre el cáncer de pulmón y el tabaquismo?". Los que quieren responder afirmativamente a esta pregunta se enfrentan regularmente con la referencia al abuelo de 80 años, que había disfrutado del humo azul desde su primera juventud, o con la referencia a una víctima de cáncer de pulmón que nunca había fumado. Obviamente, influyen muchas condiciones. Antes de que entren en juego aquí los errores de la percepción humana y la heurística, unidos a una cierta incomprensión o ignorancia de los resultados estadísticos, debemos tener claro qué constituye una *respuesta científica legítima* a una pregunta concreta que no entre en la categoría de "anécdota", "opinión" o "conocimiento a medias" o "fake news". En el caso de la relación entre el tabaquismo y el cáncer de pulmón, tenemos varias opciones como la *causalidad*, la *estadística* o el *principio lógico de exclusión*. Veámoslo a modo de ejemplo:

12.2.1 Causalidad

Nos centramos en demostrar una relación causal a nivel biológico. Por razones éticas, la experimentación con animales y seres humanos está descartada, pero puede haber estudios a nivel de células y tejidos celulares. Si podemos demostrar la causalidad de forma sustantiva y fiable, por ejemplo, que las sustancias contenidas en los cigarrillos provocan o favorecen fuertemente el cáncer, entonces no necesitamos necesariamente los siguientes casos.

12.2.2 Estadística

Se aplica la ley de los grandes números o la estadística de muestras pequeñas. Para una cuestión tan delicada y relevante desde el punto de vista político y económico, una muestra más grande es definitivamente esencial. Aquí hay que tener en cuenta posibles variables de terceros y definir la muestra con precisión. Los casos individuales son relevantes aquí sólo en la medida en que puedan demostrar causalmente que sus

mecanismos se aplican por igual a todas las personas estudiadas. Puede ser que el abuelo fumador haya fumado efectivamente durante décadas y no tenga cáncer de pulmón. Pero eso no significa que su constitución se aplique exactamente a todas las demás personas. Tal vez sea simplemente una excepción robusta y el cáncer de pulmón llegue más tarde y sea cuando ya haya muerto de otra cosa. Esto no invalidaría la tesis de que "fumar está directamente relacionado con el cáncer de pulmón", pero podría introducir variables moderadoras como los genes o similares. En general, esto hace que la situación sea más compleja, pero como resultado, se pueden distinguir posibles tipos de posibles procesos. Los análisis estadísticos complejos pueden aislar dichos factores mediadores y separarlos numéricamente de los efectos globales.

12.2.3 Principio de exclusión lógica

El hecho de que existan diferentes formas de contraer cáncer de pulmón (por ejemplo, trabajar con sustancias y polvo peligrosos) no significa que no exista una relación directa o incluso causal entre el tabaquismo y el cáncer de pulmón. Sin embargo, como científicos serios también debemos investigar los factores de protección con respecto al cáncer de pulmón o la hipótesis de si es posible fumar durante mucho tiempo sin efectos secundarios perjudiciales. También debemos averiguar exactamente qué sustancias causan el cáncer de pulmón. También estudiaremos la cuestión del potencial adictivo de la nicotina, ¡que es altísimo! - e intentamos integrar todos estos factores de influencia en un modelo complejo. La lógica en sí misma nos ayuda sobre todo a refutar las anécdotas anteriores o a convertirlas en objeto de una investigación específica. Así, podríamos obducir al abuelo después de su muerte, hacer un análisis genético, reconstruir su historia vital y su comportamiento, etc. y comparar los resultados con otros casos que murieron de cáncer de pulmón a una edad muy temprana o que nunca fumaron.

Lo que podemos hacer, entonces, es extraer categorías claras para cada una de estas variantes de estudio con respecto a la conexión entre "fumar y el cáncer de pulmón", que representan las estructuras condicionales empíricas de una multitud de resultados de estudios de los que disponemos. A continuación, podemos examinarlas lógicamente con más detalle con el *criterio conexión "cáncer de pulmón VERDADERO vs. falso"* y en la *reducción* a la esencia obtenemos diferentes tipos de cuándo existe o no la conexión pretendida y qué combinaciones de factores influyentes están contenidas en cada caso. Sin embargo, no hemos priorizado (todavía) los tipos de estudios y sus condiciones. Pero deberíamos hacerlo, como muestra el siguiente modelo de pensamiento. Es importante siempre no dejar el sentido común fuera de la ecuación. Si encontramos estrictamente reproducible experimentalmente el supuesto vínculo entre el tabaquismo y el cáncer de pulmón a nivel biológico/bioquímico, ya no necesitamos un metaanálisis. El experimento reproducible es prueba suficiente, especialmente si el modelo utilizado se demuestra empíricamente una y otra vez. Entonces ya no necesitamos preguntar a la gente cómo se siente o qué piensa.

Básicamente, deberíamos evaluar de antemano la seriedad de un estudio antes de incluirlo en un análisis comparativo. En el caso del tabaquismo, tenemos que asumir que algunos o incluso muchos de los estudios en cuestión fueron financiados por la industria tabaquera. En general, esto aumenta la probabilidad de que los estudios críticos no se publiquen en primer lugar. En consecuencia, los resultados de muchos estudios deben evaluarse con cautela o no tienen cabida en el análisis. Estas consideraciones deberían mostrar que la selección de factores condicionales para la comparación de casos individuales no es trivial. Así que podemos recopilar todos estudios patrocinados por la industria tabaquera y realizar un análisis separado para ellos. Esto podría mostrar si existe un fuerte sesgo o no.

En comparación con las explicaciones biológicas anteriores del ejemplo "fumar", que son suficientemente complejas, el ámbito social a veces va un paso más allá en lo que respecta a la complejidad de los casos individuales. Especialmente en el estudio de la experiencia y el comportamiento sociales, no se debe esperar encontrar una única causa específica para un fenómeno concreto. Es mucho más importante descubrir las constelaciones de condiciones en las que se produce el fenómeno crítico. Para ello, hay que comparar muchos casos en los que este fenómeno se muestra y esto implica la introducción de cierta imprecisión. Sin embargo, ésta no es "perjudicial" ni "obstructiva", sino que corresponde a la realidad.

Ragin (1987) desarrolló un método comparativo explícito aplicando el álgebra de Boole a datos cualitativos. El término *análisis comparativo cualitativo* (ACQ) se utiliza a menudo en la literatura. Dado que

el objetivo del análisis es reducir las condiciones iniciales a configuraciones mínimas de condiciones, denominadas *implicantes*, utilizaremos el nombre *análisis de implicantes*, que es sinónimo de ACQ. El procedimiento se basa en el *algoritmo de minimización lógica* de *Quine-McClusky* (McDermott, 1985). Este algoritmo *minimiza una función booleana*. Un importante campo de aplicación de la minimización booleana es el desarrollo de hardware en el ámbito informático. Por ejemplo, gracias a los resultados del análisis de la minimización booleana, se necesitan menos conexiones en un chip para poder realizar completamente un conjunto definido de operaciones lógicas. Esto hace que el chip sea más barato de fabricar, presumiblemente consuma menos energía, los caminos sean quizá más cortos y el chip, por tanto, más rápido, etc. Se trata, pues, de *eficiencia*, de explicar un determinado conjunto definido de resultados con un conjunto mínimo de factores condicionales. Trasladado a nuestro ámbito social, este enfoque se utiliza para encontrar en todos los casos individuales examinados (pueden ser personas, textos, pero también estudios enteros) las condiciones mínimas según Ragin (1987, p.121). Debido a su flexibilidad, el algoritmo es adecuado para el metaanálisis. Generalmente, el enfoque es apto

- para comparar un gran número de casos individuales
- para captar vínculos complejos entre condiciones,
- si se desea, para elaborar reconstrucciones parsimoniosas,
- para examinar casos individuales tanto en partes relevantes como en su conjunto (reducción temática), y
- para comparar reconstrucciones competidoras entre sí.

Estas configuraciones de condiciones tienen la cualidad de tipos. Las configuraciones de condiciones resultantes corresponden a las características que constituyen el tipo respectivo. Éstas, a su vez, se refieren a casos individuales concretos existentes. Esto significa, sin embargo, que no todas las condiciones que entran en el análisis forman parte necesariamente de un tipo posteriormente. El procedimiento elimina aquellas condiciones que no influyen en el criterio a nivel lógico. Es importante comprender que este proceso de reducción no es de naturaleza estadística o relacionada con el contenido, sino el resultado de vínculos puramente lógicos.

12.3 La formación de tipos como principio de comparación mediante la minimización lógica

A diferencia de los enfoques orientados a las variables, que parten del supuesto básico de la *agregación aditiva* de las variables individuales, la comparación orientada a los casos se basa en (Ragin, 1987, p. 51s.) para

- elaborar invariancias o conexiones constantes de significado mediante comparaciones cuidadosas de casos individuales,
- prestar más atención a la variabilidad de las configuraciones condicionales significativas que a las meras distribuciones de la altivez de casos típicos, de lo que se deduce que incluso un único caso contradictorio debe tenerse en cuenta.
- captar los casos como conjuntos, es decir, ver las condiciones del caso en interdependencia, lo que constituye este caso particular, pero no comprender las condiciones individuales en función de una distribución de población, así como
- examinar con precisión cómo las condiciones globales inferidas conducen a conclusiones distintas en contextos y configuraciones diferentes.

Para aplicar las reglas del álgebra de Boole a la interpretación de los datos, necesitamos una condición inicial definida en la que debe existir la información. En el caso de la minimización booleana, se trata de una tabla binaria o de valores de verdad. Para ello, se reducen radicalmente los significados encontrados en cada

fichero a valores de verdad, es decir, a un *sistema binario* compuesto únicamente por [VERDADERO, falso], [1, 0] o [sí, no]. Así, uno se contenta con la notación binaria *condición es VERDADERO* (= dado) o *condición es falso* (= no dado). No importa si se trata de condiciones genuinamente cualitativas, es decir, de interpretaciones y la asignación de categorías adecuadas, o de valores medidos de una característica cuantitativa. Podríamos transformar fácilmente los principales resultados de estudios enteros en tales valores de verdad para llevar a cabo un metaanálisis lógico.

Veamos un *ejemplo ficticio de investigación empírica educativa*: En búsqueda de las condiciones del *éxito escolar (E)* como fenómeno crítico, n = 38 estudios se centraron en k = 3 factores:

- *Calidad de la enseñanza (U)*
- *Aptitud de los alumnos (B)*
- *Tamaño de la clase (K)*

Los datos proceden de diferentes perspectivas y, por tanto, integran tipos de datos cuantitativos y cualitativos a nivel lógico.

- Los valores de las pruebas (valores medios de las clases) están disponibles para el éxito escolar E y la aptitud individual B, es decir, datos numéricos.
- El tamaño de la clase K está disponible como valor de frecuencia, también datos numéricos.
- La calidad de la enseñanza U es estimada por expertos a partir de datos de observación. Para la calidad de la enseñanza impartida, los investigadores disponen de datos cualitativos-interpretativos que han sido codificados dentro del paradigma de codificación.

Todos estos datos se convierten ahora en *valores de verdad*, por ejemplo, utilizando *la mediana* como criterio de corte. En concreto, esto significa que el 50 % de los portadores de datos por debajo de la mediana reciben aproximadamente un 0 y el otro 50 % por encima de la mediana reciben un 1. De este modo, el conjunto de portadores 0 y 1 está equilibrado, ya que la mediana divide una distribución en el 50 % del área. Ahora podríamos discutir ampliamente si los valores que corresponden exactamente a la mediana reciben un 1 o un 0. Esta decisión debe tomarse a lo largo de los datos, ya que podría introducir un *pequeño sesgo*, siempre que haya valores que tomen exactamente la mediana. Del mismo modo razones de contenido pueden aconsejar que el punto de corte se fije en un lugar diferente. La desviación del criterio de la mediana debe estar siempre bien justificada desde el punto de vista del contenido. Por tanto, en el análisis posterior se trabaja con los valores *éxito escolar* E VERDADERO/FALSO [1, 0], *calidad de la enseñanza* U VERDADERO/FALSO [1, 0], *aptitud* B VERDADERO/FALSO [1, 0] y *tamaño de la clase* K VERDADERO /FALSO, es decir, *pequeña/grande* [1, 0]. Supongamos que en los 38 casos disponibles (es decir, clases) se notan ocho constelaciones de condiciones (véase la Tabla 12.3). Hemos elegido la notación [1, 0].

En principio, la minimización lógica procede de tal manera que todas las constelaciones que contienen el *éxito escolar como E VERDADERO* se comparan entre sí. Si, en casos por lo demás similares, *una única condición* se da como VERDADERA en algunos casos y como FALSA en otros, lógicamente se puede descartar como *irrelevante porque no está conectada coherentemente con el criterio*. En palabras de Ragin (1987), si dos expresiones booleanas difieren sólo en una única condición causal pero conducen al mismo resultado, esa condición causal que las distingue es redundante, es decir, no tiene ningún efecto y, por tanto, es irrelevante. Por lo tanto, se le puede eliminar, dando lugar a una expresión más simple y corta. Esta forma de reducción de datos se equipara a la lógica del diseño experimental.

Técnicamente, esto puede expresarse aún más claramente en el lenguaje de la lógica formal:

Recordatorio 12.3: Minimización booleana

Si dos términos de conjunción unidos por disyunción sólo difieren por la negación de una única variable, estos dos términos pueden fusionarse, eliminando la variable en cuestión (algoritmo Quine-MyCluskey).

Tabla 12.3: *Estudio ficticio del éxito escolar (tabla de valores de verdad)*

Constela-ción	*Éxito escolar (E)*	*Calidad de enseñza (U)*	*Aptitud de alumnos (B)*	*Tamaño* Clase*(K)*	*Tamaño (de casos)*	Caso
1	1	1	0	1	6	1
2	1	0	1	0	5	1
3	1	1	1	0	2	0
4	1	1	1	1	3	0
5	0	1	0	0	9	1
6	0	0	0	1	6	1
7	0	0	1	1	3	0
8	0	0	0	0	4	0

El algoritmo funciona siempre hasta que ya no es posible reducir más. Para ello, compara todas las combinaciones de condiciones entre sí. No es posible un atajo, por lo que en el caso de muchas condiciones (configuraciones de condiciones complejas) tiene lugar un número correspondientemente grande de comparaciones. Si se dispone para el análisis de n = 20 configuraciones de condiciones iniciales diferentes a través de los casos, entonces en la primera pasada, es decir, en la primera minimización lógica, (20 - 1) + ... + (20 -19) = 190 comparaciones deben llevarse a cabo, es decir

$$\sum_{n-(n-1)}^{n-1}$$

comparaciones, es decir, sum(1:(20 -1)). La tabla 12.4 muestra, a modo de ejemplo, las comparaciones para cuatro condiciones. El número absoluto de comparaciones adicionales en las siguientes rondas depende de cuánta redundancia reduzca el algoritmo y, en consecuencia, de cuántas comparaciones sigan siendo factibles. De todos modos, desaconsejamos las comparaciones demasiado complejas, ya que suelen ser demasiado específicas, es decir, no lo suficientemente generales. Además, apenas pueden interpretarse a la vista de los numerosos detalles, por lo que prácticamente se pasa por alto *el hilo conductor* entre los casos y a través del material de datos. Aquí *menos es más*, porque se trata de lo esencial, no de los detalles. Teóricamente, podríamos pensar combinatoriamente sobre qué posibles configuraciones de condiciones posibles y cuáles de ellas se dan empíricamente para hacernos una idea de la variabilidad. Sin embargo, en la mayoría de los casos esto nos llevaría a confundir montañas de datos y probablemente no exista un nivel de comparación para discutir las configuraciones empíricas de condiciones con las teóricamente posibles.

La regla general es que *la calidad requiere una cantidad mínima* para ser eficaz. En concreto, nos preguntamos por la relación entre el número de afecciones y el número de casos. Dentro de una amplia tolerancia, esto puede ser causa de interpretaciones erróneas, concretamente siempre que el número de casos sea inferior a la cantidad de condiciones examinadas. Si son muchas las condiciones y pocos los casos, o incluso menos casos que condiciones, desaconsejamos utilizar el análisis, sino reducir primero las condiciones. Esto significa a menudo aumentar el nivel de abstracción hacia una mayor generalizabilidad. Por ejemplo, conocemos un caso en el que se analizaron 12 condiciones sobre 12 casos y lo que "salió" fueron supuestamente 12 "tipos" diferentes. El excesivo nivel de detalle ocultó por completo los puntos en común entre los casos, por lo que el procedimiento no produjo ninguna ganancia de conocimiento. Más bien se creó la ilusión de que existían tipos completamente diferentes y fue necesario convencer a muchos de que

esos 12 tipos no eran un resultado que hubiera que interpretar, sino artefactos si se quería hacer una comparación generalizada. El procedimiento obtiene su *elegancia de la sencillez* para descubrir combinaciones claramente separables, no preservar la unicidad con todos los detalles y, al mismo tiempo, creer poder seguir identificar patrones de casos cruzados. Un modo de movimiento perpetuo no existe en este procedimiento.

Tabla 12.4: *Minimización booleana (comparaciones posibles para cuatro condiciones)*

no.	Comparación de condiciones
1	k_1 vs. k_2
2	k_1 vs. k_3
3	k_1 vs. k_4
4	k_2 vs. k_3
5	k_2 vs. k_4
6	k_3 vs. k_4

Veamos el resultado de la minimización lógica para el ejemplo anterior Éxito escolar E VERDADERO para las cuatro primeras constelaciones de condiciones. Para una mejor comprensión, indicamos los valores de verdad con [1, 0]-valores y con *letras mayúsculas (= VERDADERO)* o *minúsculas (= falso)*. La tabla 12.5 contiene las comparaciones pertinentes de las constelaciones de condiciones que conducen a reducciones. Marcamos cada condición eliminada con un guión. Con la minimización booleana, todas las condiciones deben compararse siempre entre sí (véase la Tabla 12.4). Podemos mostrar las posibles comparaciones de configuración en R (`ptIV_qual_Boole_logical-minimization.r`). Para ello necesitamos el paquete de R `combinat`.

```
> # combinations comparisons Boolean minimization
> noquote(t(combn(paste("k",1:4,sep=""),2)))
[,1] [,2]
[1,] k1 k2
[2,] k1 k3
[3,] k1 k4
[4,] k2 k3
[5,] k2 k4
[6,] k3 k4
```

Podemos implementar fácilmente una comparación de reducción singular única en R. Para todo el conjunto de datos, utilizaremos más adelante funciones del paquete QCA de R. Éstas son más eficientes y reducen tablas enteras. Nuestro ejemplo aquí sólo sirve para demostrar el procedimiento. En primer lugar creamos dos vectores con los valores de verdad para k1 y k4 y nombramos los elementos según su origen con U, B y K.

```
# minimal example logic minimization
k1 <- c(T,F,T)
k4 <- c(T,T,T)
names(k1) <- names(k4) <- c("U","B","K")
```

Tabla 12.5: *Minimización booleana (comparación de constelaciones de condiciones)*

Comp.	Notación						Interpretación
	Letras			binario			
k_1 vs. k_4	U	b	K	1	0	1	La aptitud es irrelevante en estas dos constelaciones de condiciones, ya que *sólo difieren en un lugar con respecto al criterio*. **U-K (1-1)** permanece, ya que suprimimos el medio **b** o **B** qua algoritmo.
	U	B	K	1	1	1	
	U	-	K	1	–	1	
k_2 vs. k_3	u	B	k	0	1	0	La calidad de la enseñanza puede eliminarse como condición en estas constelaciones. Lo que queda es **-Bk (-10)**.
	U	B	k	1	1	0	
	-	B	k	–	1	0	
k_3 vs. k_4	U	B	k	1	1	0	En estas constelaciones, el tamaño de la clase no afecta al éxito escolar. **UB- (11-)** se mantiene.
	U	B	K	1	1	1	
	U	B	-	1	1	–	

Tabla 12.6: *Minimización booleana (comparación de una constelación de condiciones)*

	Constelaciones			Abbrev.	Nota
	U	B	K		
k_1	1	0	1	UbK	
k_4	1	1	1	UBK	
k_1 vs. k_4	=	≠	=	UbK vs. UBK	Si dos condiciones fueran diferentes, no estaría claro y no se podrían omitir sin más, es decir, no sería posible la minimización.
Acción	+	–	+		Las condiciones idénticas se mantienen con la condición de que sólo una condición sea diferente en conjunto en la comparación.
Resultado	U	-	K	U-K	

A continuación enlazamos ambos vectores lógicamente y examinamos en cuántos y en qué lugares difieren.

```
> k1 & k4
U B K
TRUE FALSE TRUE
> red.l <- length(id <- which(!(k1 & k4)))
> id
B
2
> red.l
[1] 1
```

Si sólo difieren en un lugar, éste puede ser eliminado por el algoritmo. Para ello utilizamos el marcador de posición "NA" (=not available/no disponible), que marca los datos que faltan en R. Esto nos permite no tener que cambiar el tipo de datos "`logica`" de los vectores.

```
> if(red.l == 1) cat(paste("reduction, because there is only one
+ difference at:\t",names(k1[id]),"\n",sep=""))
reduction, because there is only one difference at: B
> k1[id] <- NA
> str(k1)
Named logi [1:3] TRUE NA TRUE
- attr(*, "names")= chr [1:3] "U" "B" "K"
> k1
  U   B   K
TRUE  NA TRUE
```

En nuestro caso, los tres tipos descritos ya no pueden compararse entre sí y, por tanto, no pueden reducirse más. Así pues, el resultado está disponible – y podemos interpretarlo de la siguiente manera.

Tabla 12.7: *Éxito escolar ficticio (criterio de resultados TRUE)*

Categ.	Valor	Lógica	Categ.	Valor	*Significación*
U	TRUE	AND	K	TRUE	buena enseñanza AND clase pequeña
B	TRUE	AND	k	falso	alta aptitud AND clase numerosa
U	TRUE	AND	B	TRUE	buena enseñanza AND alta aptitud

Tabla 12.8: *Estudio Marcelo (1991, entrevistas con profesores principiantes)*

Categoría	Contenido/Código	Nota
A	sí mismo	
B	Relación profesor-alumnos	
C	Métodos de enseñanza	
D	Problemas de disciplina	CRITERIO
E	Motivación de los alumnos	
F	Clima social del aula	

La tabla 12.6 muestra el procedimiento en detalle para el primer ejemplo (k1 frente a k4). Si resumimos los tres resultados de las tres comparaciones, resultan tres tipos diferentes de situaciones o combinaciones de condiciones para el éxito escolar E. Para una mejor comprensión de los números binarios, volvemos a expresar los valores lógicamente VERDADEROS con *letras mayúsculas* y los valores lógicamente falsos *con minúsculas*

$$\text{Éxito escolar E} = UK + Bk + UB \qquad (12.1)$$

Pongamos el resultado en palabras: en la comparación de los estudios, el éxito escolar E se observó en los siguientes tipos de situaciones combinadas, que se resumen en la tabla 12.7. En nuestro ejemplo ficticio, se utilizó el procedimiento de minimización lógica para comparar varios estudios, es decir, para realizar un metaanálisis en el que se incluyeron originalmente datos cuantitativos y cualitativos. Así es como se puede aplicar con éxito una combinación de métodos. En el ámbito de la investigación puramente cualitativa, el enfoque de minimización lógica puede hacer realidad muchas características de calidad: Simplificación, transparencia, fiabilidad, reproducibilidad y documentabilidad.

Como ejemplo empírico de la aplicación de la minimización booleana en un estudio de entrevistas con muchos individuos, nos remitimos a un estudio español de Marcelo (1991) sobre las *experiencias de profesores principiantes*. El autor descubrió en sus entrevistas que muchos de los jóvenes profesores hablaban muy a menudo de problemas de disciplina en sus clases, pero ni mucho menos todos mencionaban este problema. En la búsqueda de diferencias críticas que pudieran explicar por qué algunos de los jóvenes profesores experimentan y expresan problemas de disciplina y otros no, el análisis se centró en seis áreas de contenido principales de las entrevistas según sus declaraciones (véase la tabla 12.8).

Tabla 12.9: *Estudio Marcelo (1991, tipos de profesores)*

Tipo	Característica
ABC	Un primer grupo de participantes en el estudio puede ser caracterizado por la configuración ABC. Este reflejan mucho sobre sí mismos, sobre las relaciones profesor- alumno y métodos de enseñanza, pero *no* sobre la motivación de los alumnos y el clima del aula.
ACEF	Un segundo grupo, que puede caracterizarse por la conguración ACEF, habla mucho de sí mismo, de los métodos de enseñanza, de la motivación de los alumnos y del clima social - pero no parece reflejar las relaciones profesor-alumno.
abcef	El tercer grupo, caracterizado por la configuración abcef, menciona con frecuencia los problemas de disciplina en la entrevista pero rara vez alguna de las otras categorías centrales.

El análisis de las conguraciones condicionales para la *condición D (= problemas de disciplina)* TRUE como *criterio* dio como resultado en el curso de la minimización booleana tres grupos distintos de constelaciones de condiciones:

$$D = ABC + ACEF + abcef \qquad (12.2)$$

Como resultado de esta reducción, aprendemos que debemos distinguir entre tres tipos de problemas disciplinarios D. La tabla 12.9 contiene la interpretación de estos tipos. Con esta información podemos ir directamente a la educación y tratar de concienciar desde el principio, o podemos entrenar a los recién llegados a la profesión y centrarnos en estas áreas problemáticas. También sería posible desarrollar una herramienta de cribado para asignar a los posibles candidatos al tipo adecuado en una fase temprana. Las posibles aplicaciones de estos resultados son muy diversas.

Por último, una minimización lógico no es sólo un procedimiento para analizar conjuntos completos de datos *en la fase final* de una investigación. Como *heurística ya contribuye a la categorías* para interpretaciones posteriores, incluso cuando sólo se han analizado unos pocos archivos de datos y la base de comparación es, por tanto, todavía modesta. Aquí es posible comprobar la coherencia interna del enfoque de interpretación con la minimización lógica. Las constelaciones de condiciones para los casos en los que el criterio es VERDADERO (problemas de disciplina D en el último ejemplo) y los casos en los que es falso (problemas de disciplina d) *deben ser completamente diferentes*. Es decir, *no puede darse una constelación idéntica en ambas condiciones del criterio*. Si no es así, se ha encontrado un claro indicador de deficiencias estructurales en el proceso de interpretación y, por tanto, en el propio sistema de codificación. Entonces el sistema de interpretación necesita una revisión sustancial. Por supuesto, es ventajoso ser consciente de tan grave inconsistencia lo antes posible en el proceso de análisis de los datos y no cuando uno quiere escribir su trabajo y "sólo comprobar rápidamente si todo es todo es coherente".

El procedimiento de minimización lógica parece *indispensable* y, de hecho a lo largo de todo el proceso de análisis – desde el examen inicial del enfoque interpretativo como se acaba de describir, hacia el final del análisis en el resumen de los resultados o la de formación de tipos, en la agrupación de hallazgos individuales, en la diferenciación de expedientes archivos (casos) o hablantes, al destacar textos clave, etc. Porque, como señala Ragin (1987), la cantidad de trabajo que suponen las comparaciones aumenta geométricamente con cada archivo adicional, ¡sino exponencialmente con cada categoría o condición adicional!

12.4 Implicantes primarias y esenciales

Volvamos primero al caso ficticio anterior del éxito escolar. Hasta ahora, siempre hemos hablado de *constelaciones o configuraciones condicionales* con referencia a un criterio. Aquí, las condiciones son resultados de datos reducidos a valores de verdad por minimización lógica. La tabla 12.10 contiene cuatro constelaciones de este tipo para el valor VERDADERO del *criterio E* (éxito), que volveremos a simbolizar con las letras iniciales en lo sucesivo para una mejor comprensión. La ecuación booleana para ello es

$$\textit{Éxito escolar } E = \text{UbK} + \text{uBk} + \text{UBk} + \text{UBK} \qquad (12.3)$$

Las condiciones de E se resumen como la suma lógica de los diferentes productos lógicos de los factores individuales. Por tanto, los signos de suma representan un OR lógico. En las condiciones individuales, es decir, en los términos algebraicos como UbK, los elementos U, b y K están unidos por la operación lógica AND. La ecuación de las constelaciones constelaciones para E encontradas empíricamente se lee así

Caso 12.1: Condiciones booleanas en abstracto

El éxito escolar E se observa en los casos
(U y b y K) o (u y B y K) o (U y B y k) o (U y B y K).

Esto nos da la ecuación empírica, aún no reducida, de todos los términos básicos relevantes. Al menos, esto ya nos indica que no debemos considerar los efectos de las condiciones individuales de forma aislada, sino necesariamente en el contexto de otras condiciones. En consecuencia, el *éxito escolar E* puede observarse ciertamente *en grandes clases escolares k*, pero sólo en el *contexto* de las condiciones *uB o UB*. Sin embargo, el objetivo de la minimización booleana no se consigue enumerando todas las configuraciones observadas en una ecuación lógica, sino sólo cuando se han *simplificado*, como se ha descrito anteriormente, las expresiones de la ecuación mediante el procedimiento de comparación lógica. Aquí es donde entra en juego el *concepto de implicantes*, que se define mediante la *teoría de conjuntos*. Una expresión booleana implica a otra, si la segunda expresión "implícita" es un subconjunto de la primera. Lo que valen en la teoría de conjuntos son el principio jerárquico y la expresividad limitada de los subconjuntos sobre el conjunto total.

En nuestro ejemplo, U incluye o implica todos los casos de UbK o UBk o UBK, es decir, estos casos son un subconjunto de todos los casos de UbK o UBk o UBK. son un subconjunto de todos los casos en los que se ha observado U. O dicho de otro modo

Caso 12.2: Condiciones booleanas formuladas

El *éxito escolar E* se observa en todas las clases en las que tiene lugar una buena enseñanza U, pero *sólo en conexión* (= AND lógico) con otras condiciones – a saber bK o Bk o BK.

Tabla 12.10: *Estudio ficticio del éxito escolar (implicados primitivos y primarios)*

		Implicantes primitivos			
		UbK	uBk	UBk	UBK
Implicantes primarios	UK	×			×
	Bk		×	×	
	UB			×	×

Combinando todos los términos booleanos originales (o primitivos) y la siguiente minimización, es decir, excluyendo las condiciones irrelevantes en determinadas combinaciones, tenemos una minimización lógica sobre la ecuación ya conocida

$$\text{Éxito escolar } E = UK + Bk + UB \tag{12.4}$$

Los términos UK, Bk y UB son los *implicantes primarios* del criterio *éxito escolar E* (VERDADERO). Siempre es importante añadir qué valor de verdad tiene el criterio, ya que más adelante pedimos configuraciones de condiciones para *la negación del criterio*, es decir, la *falta de éxito escolar*. El análisis y la comparación de ambas formas del criterio de verdad es esencial en el análisis de implicantes. Según los resultados empíricos, la ecuación original de cuatro miembros se redujo a una de tres por minimización. Ahora, sin embargo, en muchas tablas de condiciones – no en todas – es posible que los implicantes primarios puedan reducirse lógicamente aún más, de modo que sólo queden los *implicantes lógicamente esenciales*, cuasi el núcleo, que entonces realmente no es accesible a ninguna reducción posterior. Para ello sirven las combinaciones tabulares según el siguiente patrón, en el que las *expresiones primitivas* definen las columnas y los *implicantes primarios* las filas (véase la Tab. 12.10).

Si ahora intentamos reducir aún más el número de implicantes primarios, de modo que tengamos el concepto de subconjunto, podemos abarcar todas las expresiones primitivas con un número mínimo de implicantes primarios. Para cubrir todas las expresiones primitivas con un número mínimo de implicantes primarios, procedemos como sigue (véase la tabla 12.11):

- Tabulación cruzada de implicantes primarios (líneas) y sus elementos individuales. Corresponde a la unión de todos los elementos de todos los implicantes primarios.
- Ahora se examinan estos elementos para ver qué elementos aparecen una sola vez en la tabulación cruzada y cuáles aparecen con más frecuencia.
- Se toman aquellos elementos primarios en los que se dan estos elementos singulares. Éstos son los implicantes primarios esenciales. Todos los demás son implicantes primarios cuasi "normales" y están ya cubiertos por los implicantes primarios esenciales.

Tabla 12.11: *Estudio ficticio del éxito escolar (implicados primarios y esenciales)*

		Elementos			
		U	B	k	K
Implicantes primarios	UK	○			×
	Bk		○	×	
	UB	○	○		
	∑	2	2	1	1

En la tabla 12.11, una cruz representa una única mención (aparición) y un círculo, al menos dos. Rastreamos la afiliación de las "cruces" hasta los implicantes primarios correspondientes. Éstos corresponden a los implicantes esenciales primarios. Para nuestro ejemplo, identificamos como implicantes esenciales sólo UK y Bk.

$$\text{Éxito escolar } E = UK + Bk \tag{12.5}$$

Esto significa que el éxito escolar E se observa con una *buena enseñanza en clases pequeñas* o con *alumnos superdotados en clases numerosas*. El implicante primario UB (buena enseñanza y alumnos superdotados) se incluye como subconjunto en los implicantes esenciales y, por tanto, no cuenta como implicante primario esencial. Aunque este conjunto de solución de implicantes esenciales es la solución mínima lógicamente correcta, no siempre parece empíricamente significativo reducir más los implicantes primarios a su esencia. Es intuitivamente claro que el implicante UB (*buena enseñanza de los alumnos superdotados*) también implica el éxito escolar E, pero UB nunca puede observarse *sin el factor K/k (tamaño de la clase)*.

Es precisamente este tipo de información, que ciertos factores ejercen sus efectos dependientes o independientes de otros, lo que no siempre resulta tan obvio. Por lo tanto, siempre se debería dejar en la minimización a los implicados primarios. Esto ayuda a la interpretación. Quienes sepan manejar los implicantes esenciales pueden y deben utilizarlos con cuidado, pero sean conscientes de que esto puede dificultar la comunicación a nivel de contenido. No todo el mundo podrá seguir la argumentación, ya que se basa en consideraciones lógico-formales y no de contenido. Consideramos que el criterio relevante es el contenido y su inserción en un contexto, y no la pura lógica formal – aunque sea técnicamente correcto.

12.5 Secuencia de pasos de la minimización lógica

Para poder utilizar la minimización booleana, hay que disponer de hipótesis procedentes de las consideraciones preliminares sobre el diseño de una investigación y/o de los análisis realizados hasta ese momento acerca de qué factores (condiciones) podrían estar relacionados con un determinado resultado crítico (véase la abstracción en la Tab. 12.12). En el ejemplo ficticio, se partió de la hipótesis de que el éxito escolar está influido por la calidad de la enseñanza, la aptitud de los alumnos y el tamaño de la clase escolar. Las hipótesis de este tipo son la condición previa general y surgen de supuestos teóricos, de los resultados de otros estudios o durante el proceso de análisis, es decir, el examen del material de datos. Al formular tales hipótesis, hay que seguir el principio de simplicidad o parsimonia (= la navaja de Occam) para mantener el número y la composición de los implicados dentro de un marco manejable que (aún) facilite interpretaciones significativas. Esto significa alcanzar un nivel adecuado de abstracción en el nivel de codificación (metacódigos, Huber y Gürtler, 2012, capítulo 7.3) y mantenerlo de forma coherente.

Tabla 12.12: *Minimización lógica (abstracción y proceso, valores de verdad)*

Caso	Criterio	Factor A	Factor B	Factor C	Factor [...]
1	VERDADERO o falso	VERDAD. o falso	VERDAD. o falso	VERDAD. o falso	...
2	...	...	...	...	...
3	...	...	...	...	...
...	...	...	...	...	...

1. En un primer paso, se seleccionan los datos cuantitativos y/o cualitativos disponibles sobre las hipotéticas influencias o factores y su presunto impacto o criterio para la aplicación de la minimización lógica. El procedimiento en sí no selecciona, es decir, la elección de las condiciones y criterios es en última instancia arbitraria, mejor dicho, exclusivamente de carácter sustantivo.
2. En un segundo paso, los códigos (o los datos de que se disponga) se convierten en valores de verdad, es decir, en los números binarios 1 ó 0. Estos datos se introducen en una tabla de valores de verdad, en la que las columnas están definidas por los factores y el criterio, y las filas por los casos individuales, como muestra la tabla 12.12. Como ya se ha mencionado, convencionalmente el corte para la asignación de los datos individuales a la categoría VERDADERO o FALSO se realiza en cada columna en el valor percentil 50, es decir, la mediana. Los criterios de corte alternativos requieren una justificación situada en el dominio del objeto. Se puede generar una tabla binaria directamente a partir de los datos iniciales en función del criterio elegido. En nuestro ejemplo ficticio, podríamos decidir que todas las clases con hasta 20 alumnos se incluyan en el análisis como pequeñas (VERDADERO), las que tengan más de 20 alumnos como grandes (falso). Pequeño se considera VERDADERO porque las clases pequeñas suelen asociarse a una mejor enseñanza. En última instancia, sin embargo, la clasificación es arbitraria y debe elegirse de tal forma que no surjan problemas de interpretación más adelante.
3. Con ayuda de la tabla de valores de verdad y del algoritmo de minimización lógica, se determinarán los implicados primarios del criterio. Con tablas de valores de verdad más grandes, especialmente con un mayor número de factores, es aconsejable utilizar programas informáticos que procesa la tabla según las reglas de la minimización booleana. El resultado de este paso es una ecuación booleana parcialmente reducida con implicantes primarios (Ragin, 1987).
4. Si se pretende llegar a la ecuación booleana mínima desde el punto de vista lógico, en un tercer paso se puede intentar reducir aún más con la ayuda de una tabla de implicantes primarios (véase más arriba) para determinar los implicantes esenciales. Dependiendo de la situación de los datos, esto no siempre tendrá éxito, porque es bastante concebible que los implicantes primarios ya representen la solución mínima.
5. El resultado puede discutirse ahora en términos de contenido.
6. Si existe un criterio positivo, es decir, se ha utilizado hasta ahora el valor de verdad *VERDADERO*, el siguiente paso debería ser llevar a cabo la minimización lógica para el mismo criterio, pero con el valor de verdad *falso*. Las razones para ello se explican a continuación.

12.6 Análisis de criterios - resultado positivo y negativo

Una cosa debe quedar clara sobre el procedimiento de minimización lógica: el procedimiento funciona *lógicamente y no en términos de contenido*. Los errores de contenido, como los causados por una elección de codificación poco sensata, un nivel de abstracción incorrecto, condiciones incoherentes entre casos, etc., no pueden ser detectados fácilmente por el propio procedimiento. Esta es una de las razones por las que siempre *insistimos en los criterios positivo y negativo* para realizar el análisis. Es decir, se comparan los casos

en los que el criterio es VERDADERO con aquellos en los que el criterio es falso. Esto permite estimar un espacio de discusión que es, en la medida de lo posible, exhaustivo en cuanto a los posibles resultados del criterio en el dominio dicotómico. Si sólo tomamos el resultado positivo o negativo del criterio sin examinar el otro punto de vista, en principio se desecha la mitad de la información. Entonces no sólo nos resulta difícil comprobar la coherencia de nuestro propio sistema de interpretación utilizando este procedimiento, sino que también tenemos problemas para responder a nuestra pregunta de investigación en todas las direcciones.

Un ejemplo ilustra este problema. Recordemos el ejemplo anterior del éxito escolar E. Las condiciones individuales eran la calidad de la enseñanza U, la aptitud B y el tamaño de la clase K. Si examinamos sólo el éxito escolar y no el fracaso, es decir, la expresión negativa del criterio éxito escolar, pasamos por alto posibles variables de influencia. En el caso del éxito, la ecuación booleana es éxito escolar E = UK + Bk + UB. Ahora bien, en el caso de fracaso, ocurrirá que no se trata simplemente de la negación de la ecuación en cuestión, como puede comprobarse empíricamente con facilidad. La causalidad es asimétrica y la explicación del resultado positivo no es la misma que la simple inversión del resultado negativo.

Lo que queda claro de este modo es que las condiciones que intervienen pueden, en principio, desempeñar un papel tanto en el caso de éxito como en el de fracaso ya que no variamos las condiciones iniciales para el análisis tanto del resultado positivo como del negativo. Sin embargo, el papel de las configuraciones individuales de las condiciones cambia en su expresión debido a los demás factores de influencia que también intervienen. Si no tuviéramos ni idea del caso, podríamos suponer que, en el estudio del caso ficticio, el fracaso escolar se produce cuando la aptitud es baja y el tamaño de la clase es grande, o cuando la calidad de la enseñanza es baja y la aptitud también es débil. Esto último podría ser independiente del tamaño de la clase. Pero todo esto no son más que conjeturas. Sólo la comprobación empírica, el análisis de los datos, aporta la esperada ganancia de conocimiento.

Si, por el contrario, elegimos tanto el resultado positivo como el negativo del criterio, el resultado es que ahora se genera un espacio de resultados uniforme que no sólo pregunta por un criterio positivo, es decir, por las condiciones del éxito escolar como en el ejemplo anterior. Consideramos que ambas cuestiones son equivalentes en el curso de este tipo de análisis y, por lo tanto, ambas deben realizarse siempre. Sólo el conocimiento de las configuraciones de las condiciones mínimas de ambas expresiones del criterio permite una evaluación exhaustiva de la gravedad de los elementos individuales de las condiciones.

A continuación implementamos esto en R – tanto para el resultado positivo como para el negativo del éxito escolar (`ptIV_qual_Boole_case_school-success.r`). Primero leemos los datos:

```
> # artificial school success example
> SE <- read.table("school-success.tab", header=TRUE, sep="\t")
> SE
  E U B K freq
1 1 1 0 1 6
2 1 0 1 0 5
3 1 1 1 0 2
4 1 1 1 1 3
5 0 1 0 0 9
6 0 0 0 1 6
7 0 0 1 1 3
8 0 0 0 0 4
> SE.ne <- SE[,c("E","U","B","K")]
```

La minimización lógica real consiste en los pasos de negar el resultado, determinar si es un caso negativo o positivo, y algún trabajo preliminar como encontrar subconjuntos con `superSubset()`. A esto le sigue la creación de una tabla de valores de verdad con `truthTable()`, que conduce directamente a la minimización lógica con `minimize()`.

```
> # outcome
> outcome <- "E"
> # positive case
> neg.out <- FALSE
> # superSubset
> SE.susu <- superSubset(SE.ne, outcome=outcome)
```

```
> print(SE.susu)
       inclN RoN   covN
----------------------------
1 U+B  1.000 0.500 0.667
2 U+~K 1.000 0.500 0.667
3 B+K  1.000 0.500 0.667
----------------------------
> # truth table
> SE.TT <- truthTable(data=SE.ne, outcome=outcome,
      neg.out=neg.out, complete=TRUE,
      show.cases=TRUE, sort.by="incl")
> print(SE.TT)
OUT: output value
n: number of cases in configuration
incl: sufficiency inclusion score
PRI: proportional reduction in inconsistency
  U B K  OUT n incl  PRI   cases
3 0 1 0   1  1 1.000 1.000 2
6 1 0 1   1  1 1.000 1.000 1
7 1 1 0   1  1 1.000 1.000 3
8 1 1 1   1  1 1.000 1.000 4
1 0 0 0   0  1 0.000 0.000 8
2 0 0 1   0  1 0.000 0.000 6
4 0 1 1   0  1 0.000 0.000 7
5 1 0 0   0  1 0.000 0.000 5
> # logic minimization
> SE.mini <- minimize(input=SE.TT, outcome=outcome,
   neg.out=neg.out, details=TRUE,
   show.cases=TRUE)
> print(SE.mini)
M1: U*K + B*~K <-> E
       inclS PRI   covS  covU  cases
----------------------------------------
1 U*K  1.000 1.000 0.500 0.500 1; 4
2 B*~K 1.000 1.000 0.500 0.500 2; 3
----------------------------------------
  M1   1.000 1.000 1.000
```

Se obtiene un resumen con `print.pis()`:

```
> print.pis(SE.mini)
###################################
### Results Boolean minimization
### Outcome: E
### Criterium: TRUE
### Conditions: U | B | K
### Output of the created minimization object:
M1: U*K + B*~K <-> E

       inclS PRI   covS  covU  cases
----------------------------------------
1 U*K  1.000 1.000 0.500 0.500 1; 4
2 B*~K 1.000 1.000 0.500 0.500 2; 3
----------------------------------------
  M1   1.000 1.000 1.000

### Fundamental base products:
[1] "~U*B*~K" "U*~B*K" "U*B*~K" "U*B*K"
simplifying assumptions (if set):
$M1
[1] U B K
<0 rows> (or 0-length row.names)
### Primary implicants (charts, no reduction):
U*B + U*K + B*~K
```

```
     3 6 7 8
U*B  - - x x
U*K  - x - x
B*~K x - x -

### Solutions:
[1] "U*K + B*~K"
### Case and primary implicants:
U*K + B*~K
   U*K B*~K
1   1   0
2   0   1
3   0   1
4   1   0
5   0   0
6   0   0
7   0   0
8   0   0
### Essential implicants:
U*K + B*~K
####################################
```

Como último paso, extraemos los implicantes (primarios, esenciales):

```
> # extract primary implicants
> paste(attr(SE.mini$PIchart,"dimnames")[[1]],collapse=" + ")
[1] "U*B + U*K + B*~K"
> # extract essential implicants
> paste(SE.mini$essential, collapse=" + ")
[1] "U*K + B*~K"
```

Creamos un diagrama de Venn (véase la Fig. 12.2) con una función de R `venn()` del paquete `venn`:

```
venn(SE.mini, ilabels=TRUE, col="blue", zcolor="pink, yellow, green",
   ellipse=TRUE, borders=FALSE, box=FALSE)
```

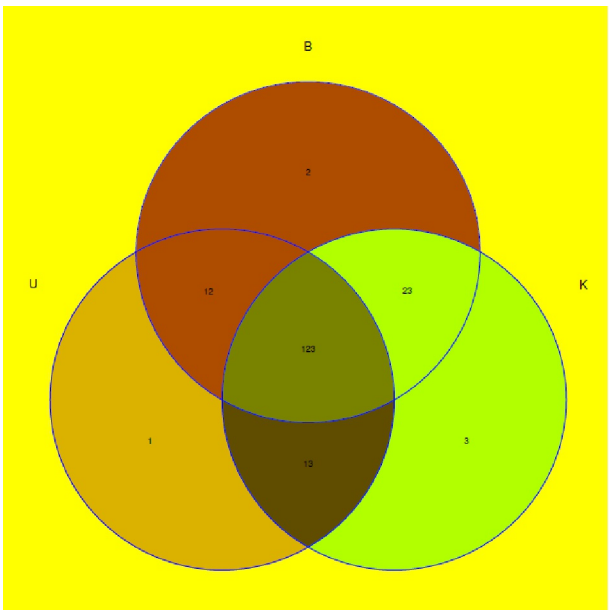


Figura 12.2 *Estudio ficticio (Diagrama de Venn)*

Ahora sigue el caso negativo. Para ello modificamos la variable `neg.out` y repetimos todos los pasos anteriores (el código R no se imprime más). Vamos directamente a los implicantes primarios y esenciales y los trazamos de nuevo. Vamos directamente a los implicantes primarios y esenciales y los trazamos de nuevo (véase la Fig. 12.2, derecha):

```
> # extract primary implicants
> paste(attr(SE.mini.NEG$PIchart,"dimnames")[[1]],collapse=" + ")
[1] "~U*~B + ~U*K + ~B*~K"
> # extract essential implicants
> paste(SE.mini.NEG$essential, collapse=" + ")
[1] "~U*K + ~B*~K"
# Venn-diagram
> venn(SE.mini.NEG, ilabels=TRUE, col="blue",
   + zcolor="darkred, yellow, green",
   + ellipse=TRUE, borders=FALSE, box=FALSE)
```

Como resultado de la minimización booleana para el caso negativo, obtenemos para los *implicantes primarios*

Fracaso escolar E = ub + uK + bk (12.6)

y para los *implicantes primarios esenciales*

Fracaso escolar E = uK + bk (12.7)

De forma análoga a la tabla 12.7, expresamos el resultado en la tabla 12.13. Vemos inmediatamente que no es simplemente una inversión de la solución de la salida positiva, sino una solución por derecho propio.

Tabla 12.13: *Éxito escolar ficticio (criterio de resultados: falso)*

Categ.	Valor	Lógica	Categ.	Valor	*Característica*
u	falso	AND	b	falso	enseñanza mala **AND** aptitud baja
u	falso	AND	K	VERDAD.	enseñanza mala **AND** clase grande
b	falso	AND	k	falso	aptitud baja **AND** clase pequeña

Así, se observa que las clases pequeñas también aparecen en el contexto del fracaso escolar, es decir, cuando van unidas a una baja aptitud. En este caso, la enseñanza en sí ya no desempeña ningún papel. En las otras dos soluciones, la mala calidad de la enseñanza desempeña un papel importante, como es de esperar intuitivamente en un contexto de baja aptitud o de clases numerosas. Por el lado de los implicados esenciales primarios, la primera solución de la tabla 12.13 se omite. Entonces tendríamos que decidir en el caso real si tampoco mencionamos la configuración distante de *mala enseñanza Y baja capacidad* en la discusión o si preferimos hacerlo. Abogamos por discutir estas configuraciones lógicamente redundantes pero sustancialmente muy significativas por todos los medios, pero manteniendo la distinción entre implicantes primarios y primarios-esenciales, introduciendo así cierta priorización en la sección de discusión y preservando así la distinción lógica.

En la figura 12.2 se muestran los diagramas de Venn resultantes. Los diagramas de Venn pueden representar gráficamente las soluciones para conguraciones sencillas.

12.7 Fuzzy logic / Lógica difusa

Además del caso dicotómico que hemos tratado hasta ahora, existe la posibilidad de trabajar a lo largo de un grado de pertenencia a una clase o categoría. Además del caso dicotómico que hemos tratado hasta ahora, existe la posibilidad de trabajar a lo largo de un grado de pertenencia a una clase o categoría. La visión binaria "afiliación a una clase sí vs. no" se suaviza entonces y se sustituye por un grado de afiliación que puede variar libremente entre 0 y 1 (véase la Fig. 12.3, ptIV_qual_Boole_fuzzy-logic.r). Entonces se puede trabajar siguiendo las mismas operaciones lógicas teóricas de conjuntos que en el caso binario (detallado en Kruse, Gebhardt & Klawonn, 1994). La imprecisión explícitamente permitida – es decir, la imprecisión relacionada con la pertenencia a una clase – se procesa lógicamente precisamente con este fin.

Esta variante del análisis implícito se denomina *lógica difusa* (*fuzzy logic*) en la literatura angloamericana y los conjuntos de datos correspondientes se denominan *conjuntos difusos* (*fuzzy sets*) y el análisis lógico correspondiente *fsQCA*. Los conjuntos de datos binarios se denominan *conjuntos crujientes* (*crisp sets*) y el análisis *csQCA* (véase la tabla 12.15). La lógica difusa no sólo se utiliza en las áreas de investigación principalmente orientadas a las ciencias sociales, que se tratan aquí. La lógica difusa se utiliza ampliamente en la ingeniería y las ciencias naturales, donde los modelos complejos a menudo sólo pueden examinarse bajo el supuesto de conceptos ideales, que, sin embargo, en la práctica no son plausibles. Especialmente en el campo de la ingeniería de control, los sistemas difusos han dado lugar a éxitos inesperados. La borrosidad resulta de la información y los modelos disponibles. Por ejemplo, si se dispone de datos estadísticos, se podría utilizar un enfoque probabilístico en lugar de uno lógico para considerar la información difusa a lo largo de métodos comunes en sus posibles márgenes de tolerancia. Si no se dispone de esos datos o si no hay un modelo de medición que describa cómo tratar la imprecisión, se puede pasar a la lógica difusa. Curiosamente, parece que los expertos pueden hacer afirmaciones intuitivas sobre estos conjuntos difusos, para que los datos estén disponibles en primer lugar. Esto es estructuralmente similar al proceso reconstrucción del conocimiento previo junto con expertos en la estadística bayesiana (O'Hagan, Buck, Daneshkhah, Eiser, Garthwaite, Jenkinson, Oakley & Rakow, 2006), si no se dispone de datos fiables de estudios previos u otras fuentes.

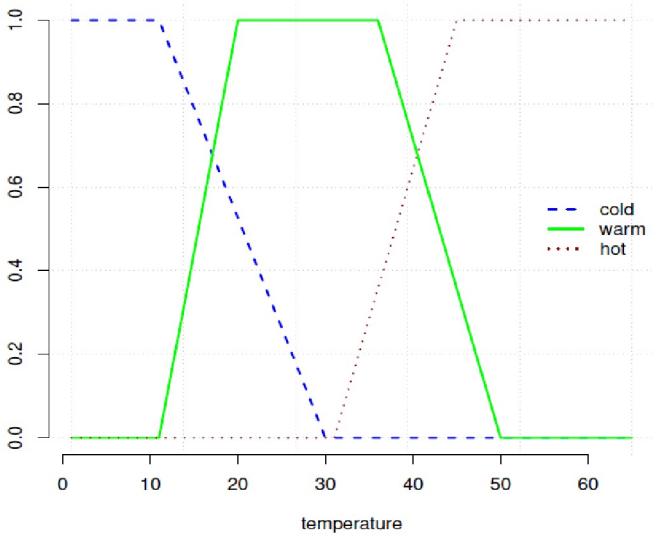


Figura 12.3 *Lógica difusa (ejemplo de medición de temperatura)*

La tabla 12.14 refleja la relación entre los operadores booleanos y la lógica difusa según los operadores Zadeh -- en honor al matemático e informático Lot A. Zadeh (1921-2017) -- con respecto a la pertenencia a agrupaciones, con los que trabaja la lógica difusa. Los sustitutos de las operaciones lógicas AND, OR y NOT, etc. son mínimo, máximo, la diferencia de uno, etc. Para VERDADERO o FALSO, la lógica difusa y las

expresiones booleanas coinciden. Otras asignaciones de las operaciones corresponden más bien a expresiones lingüísticas como *muy* o *un poco*. Son posibles diferentes asignaciones, de modo que la lógica difusa no es una aplicación estrictamente prescrita de los operadores, sino que es flexiblemente adaptable.

Tabla 12.14 *Relación de operadores booleanas y fuzzy logic (operadores de Zadeh)*

Boolean	Fuzzy Logic
`AND(x, y)`	`min(x, y)`
`OR(x, y)`	`max(x, y)`
`XOR(x, y)`	`x+y-2*min(x, y)`
`IMPLIES(x, y)`	`1-min(x, 1-y)`
`NAND(x, y)`	`1-min(x, y)`
`NOR(x, y)`	`1-max(x, y)`
`NXR(x, y)`	`1-x-y+2*min(x, y)`
`NOT IMPLIES(x, y)`	`min(x, 1-y)`
`NOT(x)`	`1-x`

Tabla 12.15 *Fuzzy Logic*

Método	Variables binarias	Pertenecia partial	Valores en el intervalo
csQCA (crisp set)	×	—si / no	[o, 1]
fsQCA (fuzzy set)	×	×	[o, 1]
mvQCA (multivalue)	—	—si / no	múltiples
gsQCA (generalized set)	—	×	múltiples

La lógica difusa permite la incorporación deliberada de la difusidad en el modelo para poder modelizar un sistema global de forma precisa a través de ella. De este modo, se presta atención a la necesidad de no definir el complejo sistema empírico con demasiada vaguedad, ya que podrían pasarse por alto características significativas. A la inversa, en la evaluación de un modelo se incluyen entonces criterios de calidad muy diferentes, que van más allá de centrarse exclusivamente en la precisión de la información, como señalan Kruse, Gebhardt y Klawonn (1994, p.2):

> "Por lo tanto, no es sorprendente que un modelo que, al incluir sistemáticamente información imperfecta, describa el sistema de interés de una forma de complejidad reducida puede ser mejor en términos de determinados criterios de calidad que un modelo que, en general, sólo permite sólo permite información precisa."

Mientras que el caso binario suele ser fácil de entender, los conjuntos difusos son mucho más complejos y difíciles de interpretar. Se utilizan conjuntos difusos cuando el fenómeno empírico investigado es difuso en sus límites y varía y/o cuando la dicotomización no es deseable. En la interpretación se debe tener en cuenta el grado de afiliación a la hora de interpretar el significado de las configuraciones de condiciones minimizadas. Además, los resultados de un análisis de lógica difusa pueden, en principio, convertirse posteriormente en un análisis binario, lo que naturalmente implica una cierta pérdida adicional de información debido a la afinación binaria de los bordes. Sin embargo, estas diferencias pueden discutirse comparativamente en términos de contenido.

Los sistemas difusos son un campo muy amplio con su correspondiente literatura técnica (por ejemplo, Ragin, 2000). En R, se pueden encontrar funciones en los paquetes `FuzzyR, FuzzyToolkitUoN, fugeR` o `lfl`, que permiten una aplicación parcialmente especializada (por ejemplo, lingüística, genética) de la lógica difusa. La minimización booleana, es decir, *fsQCA*, es posible con QCA o `QCApro`.

12.8 Valor añadido mediante el análisis de implicantes

Otra variante del análisis de implicantes es cuando se asigna más de un valor (número natural) a una condición, es decir, es de naturaleza multicategórica. Esto se conoce comúnmente como *QCA multivalor (mvQCA)* (Haesebrouck, 2016). Mientras que el caso binario indexa la presencia o ausencia de una única condición y la lógica difusa se limita a refinarla con valores entre 0 y 1, el *mvQCA* amplía la situación inicial en el sentido de que permite la presencia de diferentes categorías dentro de una condición. Así, la condición "la leche" podría subdividirse en las múltiples categorías *Vaca* (ganadería masiva), *Ecológica* (ganadería de pastos), *Soja* o *Cabra* y se codificaría en consecuencia con los números del uno al cuatro. El resto de operaciones lógicas o el procedimiento del método no cambian. Para la creación de tablas de valores de verdad se aplica, siempre que las filas contengan los casos y las columnas contengan las condiciones y el criterio:

- Los casos se agrupan en combinaciones no redundantes de condiciones
- Los valores cero y uno indican la ausencia o presencia de una condición y de forma idéntica para todos los casos. Así pues, existe una combinación de condiciones sucinta para el fenómeno investigado, dependiente del interés en el resultado positivo o negativo del criterio.
- Un caso puede ser VERDADERO o falso con respecto al criterio, pero no ambas cosas a la vez. Los datos se examinarán para garantizar que el mismo caso no se da de forma redundante y contradictoria (es decir, VERDADERO y falso simultáneamente) en una configuración de este tipo. Del mismo modo, se comprobará que en las configuraciones de condiciones resultantes ocurre lo mismo, es decir, que no existe una misma configuración de condiciones para VERDADERO y falso al mismo tiempo. Si éste es el caso, se deben analizar los datos originales muy cuidadosamente para comprobar su coherencia.

Haesebrouck (2016) ofrece un análisis detallado con ejemplos de *mvQCA* en comparación con *csQCA* y *fsQCA*. Por supuesto, uno podría preguntarse si *mvQCA* no es simplemente una forma más eficiente de *csQCA*, ya que permite expresar más categorías con menos combinaciones de condiciones debido a la asignación múltiple de categorías. El precio de esto es entonces sólo la asignación múltiple de categorías. Así que también podríamos formular el ejemplo anterior de la leche con cuatro condiciones independientes, cada una codificada en binario. En resumen, cada tabla de valores de verdad de un *mvQCA* puede, en principio, convertirse en una *csQCA* (véase la tabla 12.16).

Tabla 12.16: *Codificación de valores de verdad (csQCA vs. mvQCA)*

Método	Combinaciones de condiciones*						Criterio
`csQCA`	Vaca (Gan. masiva)	Vaca (pastos)	`Soja`	Cabra	Origen	...	Sabor
Ejemplo	0	1	0	0	1	...	1
`mvQCA`	Leche				Origen	...	Sabor
Ejemplo	2				1	...	1

*Códigos para *mvQCA*:
1 = Vaca (Gan. masiva), 2 = Vaca (pastos), 3 = Soja, 4 = Ziege

Surgen problemas en la interpretación de las configuraciones condicionales reducidas si un QCA contiene fundamentalmente combinaciones de condiciones que se contradicen entre sí. Esta situación resulta según Herrmann y Cronqvist (2009) como consecuencia de la combinación del número de casos y la dicotomización. Cuando los casos se contradicen entre sí o son ambiguos (Ragin, 1987, p.113.), es posible pasar al nivel del análisis de casos individuales para examinar con más detalle las circunstancias de los casos problemáticos. Esto puede dar lugar a una revisión de las condiciones que entran en el análisis. Ragin (ibid.) da ejemplos del procedimiento concreto. Sin embargo, al aumentar el tamaño de la muestra, cabe esperar un crecimiento cuasi exponencial del número de variables, por lo que conviene analizar muestras de tamaño medio para que todo el proceso resulte manejable. También hay que preguntarse cuánta información original se debe incluir en el análisis para lograr un compromiso entre la reducción de la complejidad y la falta de ambigüedad de los resultados. En el caso del *mvQCA*, según Herrmann y Cronqvist (2009), se dispone de un procedimiento que puede utilizarse para muestras de tamaño medio preservando al mismo tiempo cierta información sobre los conglomerados. La interpretación de los resultados es sin duda más complicada que en el caso binario, pero permite ir más allá de los límites de las categorías rígidas, lo que puede aportar su propio valor añadido, es decir, una ganancia de conocimiento en sí mismo. En el plano lógico, existen alternativas metodológicas para tales casos y preguntas, como el *análisis de coincidencias* (Baumgartner & Epple, 2013). Este método también se basa en el álgebra de Boole y es supuestamente adecuado para el análisis causal. El procedimiento puede describirse como sigue:

> „Basically, CNA eliminates factors from combinations if the resulting, reduced, combination does not occur in combination with the outcome's absence. A condition is thus removed from a combination if the latter is sufficient without this condition." (Haesebrouck, 2016, parr.. 26).

Una aplicación habitual de *mvQCA* es la cartografía de condiciones ordinales. Haesebrouck (2016) da el ejemplo de un semáforo con las características rojo, naranja/amarillo y verde. Nuestro ejemplo de la leche aquí no es de naturaleza ordinal, sino multinomial. El ejemplo del semáforo podría convertirse fácilmente (véase la tabla 12.16) en una tabla binaria más grande. La eficacia del argumento, debido a la codificación múltiple, no puede descartarse de plano. No obstante, deberíamos considerar si el caso ordinal y, del mismo modo, el grado de afiliación en lógica difusa, el pensamiento cuantitativo se reintroduce de nuevo en el modelo.

En el curso de un análisis, deberíamos preguntarnos no sólo una vez si nos gustaría responder mejor a nuestra pregunta mediante un procedimiento lógico o si sería mejor utilizar un modelo estadístico. Por ejemplo, hay artículos que tratan el análisis de implicantes como sustituto del análisis de regresión (Seawright, 2005). En nuestra opinión, estas discusiones son importantes para volver a centrarse en la propia pregunta de investigación. Obviamente, el análisis de implicantes responde a una pregunta muy diferente a la del análisis de regresión. El análisis de implicantes investiga y encuentra configuraciones condicionales mínimas suficientes para la ocurrencia de un fenómeno o para su no ocurrencia. Esto se observa independientemente de que los datos se manejen ahora de forma binaria, multinomial o con y sin imprecisión en el trazado de los límites (función de pertenencia) de las condiciones iniciales que hay que minimizar. La regresión examina la contribución proporcional de los predictores en la comparación interna, así como las posibles interacciones entre predictores y, en general, la significación práctica, es decir, el impacto que tiene el modelo global estimado o los predictores individuales con respecto a un criterio sobre una base estadística. En un proceso de varios pasos, se construye un modelo estadístico que debería ser superior a otros modelos posibles, integra casos especiales y es suficientemente complejo, es decir, el modelo ideal. En lugar de discutir en general si un método es mejor, parece tener más sentido preguntarse en qué condiciones y *para qué cuestiones* puede tener sus ventajas sobre el otro. Además, si el tipo de la información resultante es adecuado para uno u otro método.

Desde el punto de vista del software, todas las variantes presentadas pueden implementarse en R. Para minimización booleana (análisis cualitativo-comparativo), los paquetes de R `QCA`, `QCApro`, `braQCA` y `QCAtools`. Para el análisis de coincidencias, se toma el paquete de R `cna` o causalChain() en el paquete R `QCA`. Se puede encontrar un ejemplo de análisis en Weise (2014), que utiliza el análisis de Baumgartner y Epple (2013). Los antecedentes del análisis de coincidencias se explica en Baumgartner y Thiem (2015). Estas

soluciones de software proporcionan mucha más información sobre el análisis, que trataremos a continuación.

12.9 Otros parámetros de análisis

Para nuestros propósitos en el contexto del análisis de datos cualitativos o metaanálisis, suele ser suficiente concentrarse en los implicantes primarios o primarios esenciales. Éstos generan suficiente información nueva para interpretar los datos originales o, en el caso de un metaestudio, para extraer conclusiones apropiadas entre los estudios. Sin embargo, los paquetes de software disponibles permiten mucho más en términos de análisis y producen información más detallada (Thiem & Dusa, 2013a, p.88, Tab. 1 y Dus, a, 2017, respectivamente). El campo de los análisis de datos QCA puede resumirse por palabras clave con las categorías *variantes de datos*, *tipos de solución* y *procedimientos y métodos*. En Thiem y Baumgartner (2016) se puede encontrar un glosario de términos relevantes para QCA. En la siguiente lista, damos pistas sobre cómo realizar los respectivos análisis en R.

12.9.1 Variantes de datos

Los datos iniciales y su relación con los valores de verdad únicos pueden variar. Esto da lugar a diferentes análisis.

- Caso binario QCA *(csQCA)* – Sólo están disponibles los valores lógicamente VERDADERO y lógicamente FALSO.
- Multi-Value QCA *(mvQCA)* – Aquí se pueden asignar varios valores para una condición.
- Fuzzy LogicQCA *(fsQCA)* – Para el *fsQCA*, el grado de afiliación a una condición (categoría) se puede especificar con valores entre 0 y 1.
- QCA generalizado *(gsQCA)* – *gsQCA* incluye *csQCA, mvQCA* y *fsQCA* como casos especiales. La base de datos son datos difusos multinivel, es decir, una integración de todas las posibilidades de los tres casos especiales anteriores (Thiem, 2012).
- Temporary QCA *(tQCA)* – En este caso complejo, se pueden incluir elementos temporales dinámicos y alternantes; y esto conduce en última instancia al QCA de series temporales (Hino, 2009).

12.9.2 Tipos de solución del QCA

Puede ser que en los datos iniciales y las tablas de valores de verdad correspondientes aparecen configuraciones de condición a cuales no se puede atribuir lógicamente el criterio VERDADERO o falso). Esta configuración se denomina *resto lógico*. Dependiendo de cómo se produzcan estos restos lógicos y de cómo se traten, se pueden derivar distintos tipos de solución de QCA. En términos sencillos, su objetivo es convertir los residuos lógicos en elementos de pleno derecho de las configuraciones condicionales, para lo cual se hacen diferentes presuposiciones sobre los residuos lógicos.

- *Complejo o conservador* – En este tipo de solución se presupone que todos los residuos lógicos de la tabla de valores de verdad del criterio son insuficientes. Esto a su vez (véase la página de ayuda para `eQMC()` en `QCApro`) es una suposición muy fuerte sobre todos los residuos lógicos, que probablemente *no* es justificable en la mayoría de los casos empíricos y por lo tanto *no* representa un estrategia conservador, sino más bien una estrategia generalmente unilateral.
- *Intermedio* – Aquí la suposición es que los residuos lógicos en la tabla de valores de verdad no son generalmente suficientes para el criterio.

- *Parsimonia* – En este caso, se parte del supuesto opuesto de que los residuos lógicos de la tabla de valores de verdad son suficientes para el criterio. De nuevo, es cuestionable que esto sea siempre cierto en todos los casos.

Como se puede ver, los tipos de solución difieren en cómo se conceptualizan y operacionalizan los residuos lógicos. Una solución concreta para los respectivos residuos lógicos debe derivarse del contenido y el contexto. No es puramente técnica. Las ideas puramente abstractas sobre suficiencias o insuficiencias no parecen muy serias. Por lo tanto, es muy posible que los supuestos ya varíen entre los elementos. Los resultados de la minimización lógica se determinan mejor con y sin restos lógicos y posteriormente se comparan entre sí para determinar las diferencias y los puntos en común en función de las decisiones tomadas (ideas modelo sobre la decisiones tomadas sobre los residuos lógicos. Con ello se puede mejorar el análisis.

12.9.3 Procedimientos y métodos

Particularmente `QCA` y `QCApro` de R ofrecen posibilidades para realizar el proceso completo desde los valores brutos a las tablas de valores de verdad hacia las tablas a la minimización lógica y salida gráfica en forma de diagramas de Venn. Hay más análisis e información para cada paso. Informan sobre la "calidad" y las consecuencias de las decisiones (por ejemplo, dónde cortar o para la dicotomización) a lo largo del proceso o simplemente facilitan el trabajo.

12.9.3.1 Supuestos simplificados

Los supuestos simplificados giran en su mayoría en torno a la forma de tratar las soluciones lógicas (véanse los tipos de solución anteriores) para encontrar una solución para el criterio en el contexto de la minimización lógica (algoritmo de Quine-McCluskey). En caso de hipótesis simplificadas contradictorias los residuos lógicos se utilizan tanto para el caso negativo como para el caso positivo del criterio. En los supuestos simplificados, un resto lógico se establece, por ejemplo, en VERDADERO (= 1) sin estar necesariamente cubierto por un implicante primario. La dirección posiblemente conduce a tablas de valores de verdad que no están saturadas; y una gran cantidad de residuos lógicos relacionados con la dimensionalidad de la tabla de valores de verdad conduce a una mayor diversidad limitada (empírica). Por lo tanto, la solución encontrada es en realidad sólo condicionalmente inequívoca en vista de muchos residuos lógicos (véase `limitedDiversity()` en el paquete R `QCApro`, el sucesor de `QCA`). Se pueden comparar los supuestos simplificados con la salida de `minimize()` en `QCA` en la sección `SA` (= nombre del elemento de la lista) del objeto resultante. Si no hay ninguno, R produce una tabla vacía, un `data.frame()`. Ejemplos con supuestos simplificados se pueden encontrar en la página de ayuda de `minimize()`. Para nuestro ejemplo de éxito escolar, la extracción de los supuestos simplificados tendría este aspecto si algo estuviera presente, que no es el caso aquí (`ptIV_qual_Boole_case_school-success.r`):

```
SE.mini$SA
SE.mini.NEG$SA
```

12.9.3.2 Calibración

La calibración describe básicamente el proceso de traducción de los datos brutos en conjuntos de condiciones y el criterio. En *fsQCA* esto se llama *fuzzycation*. Las variables continuas pueden ser transformadas por funciones continuas. Por lo tanto, la categorización no es necesario en *fsQCA*. Thiem y Dusa ofrecen un ejemplo (2013a, p.89). La función `calibrate()` del paquete `QCA` de R realiza este proceso de traducción. Si faltan valores umbral, pueden pasarse explícitamente a la función o buscarse y determinarse mediante el análisis jerárquico de análisis de conglomerados para encontrarlos y determinarlos. En última instancia, esto corresponde a una lógica cuantitativo-estadística en la determinación de los valores umbral.

12.9.3.3 Prueba de necesidad

Supongamos que un acontecimiento A (= condición) es *necesario* para un acontecimiento B (= criterio) y viceversa, B es *suficiente* para A. A continuación, en todos los casos se puede examinar en qué medida los elementos de A y B están relacionados con su pertenencia a B. A partir de esto, se puede calcular un *coeficiente de cobertura* (cobertura de necesidad), que hace afirmaciones sobre la medida en que B se da en relación con A. El *coeficiente de inclusión* (necessity inclusion) proporciona información sobre la medida en que A es necesario para B. En el paquete `QCA` de R, esto es posible gracias a la función `superSubset()` o `pof()`.

12.9.3.4 Prueba de suficiencia

Esta prueba se deriva perfectamente de la prueba de necesidad anterior (mismas condiciones iniciales, véase más arriba) y está diseñada para complementarla. En el caso de la inclusión sucesiva, se pregunta hasta qué punto los elementos de A y B están en relación con la afiliación a A. Si el coeficiente de suficiencia es alto, se evalúa como coherente con la hipótesis "A es suficiente para B". El análisis clásico es la tabla de valores de verdad, que contiene todas las configuraciones condicionales y sus correspondientes resultados de criterio correspondientes. Si la hipótesis de éxito no puede responderse en ninguna dirección, hablamos de casos contradictorios. Esto se produce cuando al menos dos elementos de una configuración condicional conducen a resultados de criterio opuestos. `truthTable()` de `QCA` o `pof()` calculan estos coeficientes. La función permite especificar valores umbral, por ejemplo, para negar residuos lógicos o para determinar la codificación específicamente en función de los criterios. Como parámetros de ajuste adicionales `truthTable()` da valores para la inclusión (1 = superior al umbral `ic1`, 0 = no superior al umbral `ic0`, C = contradicción/inconsistencia, es decir, no superior a `ic1`, pero superior a `ic0`, véase también `?truthTable`) y `PRI` (*proportional reduction in inconsistency* / *reducción proporcional de la incoherencia*). El índice `PRI` fue propuesto por Ragin para calcular en qué medida un *mintérmino* es sucinto tanto para el caso positivo como para el caso negativo del criterio. Un mintérmino es un producto booleano, es decir, una expresión en la que las variables están todas unidas entre sí por un AND lógico. Por el contrario, los *maxterms* son expresiones booleanas, en las que todas las variables están unidas por un OR lógico. Ambas formas pueden transformarse mutuamente. La tabla de valores de verdad creada con `truthtable()` es el punto de partida para el siguiente paso, la minimización lógica con `minimize()` en `QCA` o `eQMC()` en `QCApro`. Se pueden encontrar ejemplos y aplicaciones en Thiem y Dus (2013b) y Schneider y Wagemann (2012), respectivamente.

12.9.3.5 Factorización

La factorización de expresiones booleanas mediante `factorize()` en `QCA` encuentra todas las combinaciones de factores comunes en una expresión booleana, que se puede expresar como una suma de productos o en forma normal disyuntiva. Si es posible y se desea `factorize()` produce resultados en forma de suma de productos. Un ejemplo de la página de ayuda sobre `factorize()` en R muestra el procedimiento (`ptIV_qual_Boole_case_school-success.r`):

```
> factorize("ac + aD + bc + bD", pos=FALSE)
F1: a(c + D) + b(c + D)
F2: c(a + b) + D(a + b)
```

y la salida como suma de productos

```
> factorize("ac + aD + bc + bD", pos=TRUE)
F1: (a + b)(c + D)
```

Por supuesto, los términos minimizados lógicamente también pueden factorizarse.

12.9.3.6 Pruebas estadísticas

No vamos a profundizar más en este tema. Es cierto que, en el caso de las pruebas estadísticas, la muestra debe ser grande para "alcanzar la significación"; ya hemos tratado en detalle en otra parte de este libro qué significado tiene esto y qué significación hay detrás. Para nuestros propósitos en el contexto de los datos cualitativos, las muestras tienden a ser pequeñas de todos modos, y una hipótesis estadística comprobable rara vez existe y requiere además una pregunta significativa. Se remite a los lectores interesados a `pof()` en QCA, que compara estadísticamente la inclusión calculada empíricamente (véase más arriba) con umbrales dados. Actualmente, sólo está implementado para el caso *csQCA* con una prueba binomial. Y nos preguntamos seriamente sobre el sentido y la ganancia de conocimiento debidos a tal "significación", ya que apuntamos al contexto puramente lógico de la minimización booleana. Y si quisiéramos trabajar estadísticamente aquí, optaríamos por la estadística de Bayes.

12.10 Causalidad

Por último, no podemos eludir la cuestión de la causalidad y el interés epistémico-heurístico. El propio Ragin (1987) considera el procedimiento como una forma de probar no sólo la causalidad lógica, sino la causalidad "real", es decir, las relaciones causa-efecto interdependientes en la realidad. Si en este punto no se da una definición clara de causalidad nuestra primera reacción ante tal afirmación sería responder "Nosotros no iríamos tan lejos" o "Creemos que eso es pasarse un poco de la raya". La pretensión de demostrar la causalidad en la realidad siempre nos parece una empresa un tanto temeraria, sobre todo cuando se trata de un procedimiento sensible al contexto. Probar la causalidad en empirismo puede requerir estudios experimentales y reproducibles, estudios a largo plazo (es decir, atención al curso temporal de las relaciones causa-efecto investigadas), múltiples cambios de perspectiva, gran complejidad, etc., a ser posible, todos ellos realizados en paralelo sobre un tema concreto. Los meta-análisis, sin embargo, pueden ciertamente entrar en el terreno que extrae los factores efectivos reales a través de diferentes perspectivas. Rohwer (2011, p.13) elabora, por tanto, qué comprensión de la causalidad está presente en el análisis de implicantes haciendo referencia al filósofo y economista John Stuart Mill (1806-1873), para quien la causa de un fenómeno reside en la suma de todas las condiciones (positivas, negativas) consideradas conjuntamente y la totalidad de las contingencias de cada descripción, cuya realización conduce al fenómeno. Según Rohwer (2011), la pretensión de causalidad del análisis comparativo cualitativo sigue a Ragin (1987) Mill en tres puntos. Estos están directamente relacionados con las funciones booleanas utilizadas:

- Las causas se conceptualizan como condiciones sucesivas.
- Las causas se entienden como combinaciones de condiciones que, en conjunto, son suficientes para la aparición del fenómeno. Es lo que Ragin (1987) denomina "causalidad coyuntural".
- Pueden bastar distintas combinaciones de condiciones para que se produzca el fenómeno (multi-causalidad).

Para comprender estos tres puntos, es necesaria la idea de los modelos. Así, incluso Ragin (1987, p.99) señala que "ni la necesidad ni la suficiencia existen independientemente de las teorías que proponen causas". Una buena teoría y, en consecuencia, un buen modelo son siempre el principio y el núcleo de cualquier análisis causal. En cuanto a la distinción entre causalidad

1. como consecuencia de un modelo funcional determinista que hace afirmaciones sobre reglas causales, respectivamente
2. como la creencia en relaciones deterministas entre causas y efectos empíricamente identificables

Rohwer (2011) señala que la primera está cubierta por el análisis de implicantes. Sin embargo, el enfoque es independiente de la segunda afirmación, que se centra estrictamente en las relaciones empíricas. No en vano, el autor utiliza la expresión "*creencia*" para el punto dos en el original. Con ello Rohwer no expresa

otra cosa que los modelos son casi con toda seguridad falsos, cuando se trata de una *verdadera descripción de la realidad*. No obstante, pueden ser muy útiles para obtener un modelo manejable de la realidad y representar adecuadamente las relaciones causa-efecto si, por lo demás, son prácticos y eficaces. No hay nada malo en tal concepción de la causalidad, ya que corresponde al entendimiento común que vivimos en un mundo de verdad relativa y que representamos en este libro.

El proceso es valioso, sin duda. Sin embargo, en relación con la creencia en la causalidad empírica la situaríamos no obstante -o precisamente por ello- en el plano heurístico. Heurístico, porque la elección de criterios y condiciones es arbitraria y puede seguir el propio interés propio o seguir estrictamente "sólo" las líneas de la pregunta de investigación. De nuevo, el procedimiento no se interesa por lo que se hace con los resultados. Una vez que haya trabajado con él, comprobará por sí mismo lo rápido que cambian los resultados cuando prevalecen otras condiciones de entrada. Del cambio (es decir, intercambio) de una condición a un criterio – porque condiciones y criterio son en principio intercambiables, una "característica" que ¡*se debe utilizar heurísticamente*! – no vamos a empezar. Heurísticamente, esto es fascinante y elegante, pero desde el punto de vista de las pruebas es desastroso cuando condiciones y criterios son intercambiables en el mismo estudio. Así, el procedimiento es muy sensible al cambio en relación a sus condiciones iniciales.

Este alto grado de flexibilidad hace que el procedimiento sea muy adecuado para la comprobación de hipótesis, para contrastar hipótesis, pero igualmente adecuado para generar hipótesis. ¿Qué forma asume ahora el papel central en una investigación y cómo se diseña ésta, es decir, generación de hipótesis frente a comprobación de hipótesis, viene determinada por la pregunta de investigación y el trabajo teórico previo. Los criterios de comprobación del interés de la investigación son tan estrictos como los de la investigación experimental o estadística o según el paradigma reconstructivo (véase el capítulo 11). En el sentido de una directriz para la acción, no es posible hacer una recomendación que sea siempre válida. Procedamos como procedamos, el resultado debe reproducirse y, en el caso de un metaanálisis, complementarse con al menos una o más perspectivas adicionales.

Por supuesto, el procedimiento por sí solo no basta para demostrar una verdadera causalidad empírica en el sentido de la segunda definición de causalidad anterior. Por cierto, esto lo veríamos exactamente igual en casi todos los estudios estadísticos y cualitativos, todos los cuales muy rara vez cumplen los requisitos para una auténtica prueba de causalidad. No lo vemos como una deficiencia del análisis de implicantes. Por el contrario, permite elaborar las relaciones causales lógicas entre variables de modo que resultan las *configuraciones condicionales (= conjuntos mínimos de variables)*. El enfoque se mueve así en paralelo a la estadística, que persigue el mismo objetivo. El método es elegante, fácil de aplicar y muy directo en su interpretación.

Por último, pasamos a un estudio de caso de la ciencia política en el que el QCA desempeña un papel muy extendido como herramienta analítica.

12.11 Casos prácticos de minimización lógica

Se puede aplicar el análisis de implicantes a cuestiones que no sólo suscitan interés académico o científico. La igualdad de género no avanza por igual en todos los países. Como primer ejemplo, hemos elegido un estudio sobre la representatividad de las mujeres en los parlamentos. Como segundo estudio, el ya citado capítulo sobre la AED (s. cap. 5), en el que se examinó la cuestión de qué características eran favorables para sobrevivir el desastre del Titanic en 1912.

12.11.1 Sobre la representatividad de las mujeres en los parlamentos

En su artículo de 2010, Mona Lena Krook (2010) examina la cuestión de hasta qué punto y en qué condiciones están representadas las mujeres en los parlamentos. Para evitar los problemas estadísticos causados por las diferentes poblaciones, opta por la minimización booleana. Krook compara los países occidentales con los del África subsahariana. Su punto de partida es la observación de que, si bien las mujeres constituyen la mitad de la población mundial, sólo representan el 18% de los parlamentarios del mundo (Unión Interparlamentaria, 2009a). Estas cifras ocultan marcadas diferencias entre las naciones. Mientras que Ruanda y Suecia tienen aproximadamente la misma representación femenina, Belice y Arabia Saudí no tienen ninguna parlamentaria (Unión Interparlamentaria, 2009b).

A continuación, el autor desarrolla cinco categorías que se consideran relevantes:

- el *sistema electoral* (proporcional, mayoritario o mixto)
- las *cuotas de género existentes* (sí, no)
- el *estatus general de la mujer* en la sociedad respectiva (operacionalizado como las posibilidades abiertas a las mujeres social y económicamente, pero difícil de encuestar. (socialdemócrata, conservador, liberal)
- los *movimientos de mujeres existentes* (clasificados como el grado de autonomía respecto al Estado y los partidos políticos, pero difícil de determinar globalmente (sí, no)
- presencia de *partidos liberales de izquierda* (ya que estos partidos tienden a dar a las mujeres mejores oportunidades dentro y con ellos, aunque la dinámica real es relativamente compleja, como informa Krook (valor numérico)

así como el criterio

- Porcentaje de *mujeres representadas en los parlamentos* (valor numérico en porcentaje).

Tomamos los datos directamente del artículo de Krook, los escribimos en una tabla .csv y la leemos con R (`ptIV_qual_Boole_case_Krook_women-in-parliament.r`).

```
krook.raw <- read.table("Krook_raw_QCA.csv", sep="\t",
   header=TRUE, check.names=FALSE)
krook.raw
```

En aras de la claridad, sólo utilizamos los datos de los estados occidentales. Si observamos los datos brutos, obtenemos la Tabla 12.17. Las abreviaturas de las condiciones introducidas anteriormente figuran entre paréntesis en la segunda fila, debajo del título de cada columna. Corresponden a las abreviaturas originales de Krook (2010), salvo el criterio, que no recibe su propia abreviatura en Krook. Veamos ahora los datos dicotomizados según Krook (véase el cuadro 12.18). Las reglas para ello se encuentran en el artículo original y no se decidieron algorítmicamente, sino caso por caso y en parte manualmente por el autor. y en parte manualmente por el autor:

```
krook.ne <- read.table("Krook_NE_QCA.csv", sep="\t",
header=TRUE, check.names=FALSE)
krook.ne
```

A efectos de demostración, podemos convertir las tablas para que sean más legibles. No todo el mundo puede intuitivamente con la lógica cero-uno.

```
#Letras
aquadstyle.tt(krook.ne)
#Valores de verdad en lugar de 1/0
aquadstyle.rev(aquadstyle.tt(krook.ne))
#reines AQUAD Format, effizient, dafür keine Kolumnentitel
aquadstyle.tt(krook.ne, pres.cnam=FALSE)
#Mehrfachumwandlung = Erhalt des Ausgangsformat
aquadstyle.rev(aquadstyle.tt(krook.ne))+0
```

Trabajamos con el paquete QCA de R. La función `superSubset()` nos permite encontrar una lista de implicantes que cumplan ciertas restricciones. Los detalles se pueden encontrar en la página de ayuda de la función. A continuación enumeramos los parámetros del conjunto, aunque correspondan a los valores por defecto.

```
# define parameters R-Code
# superSubset
outcome <- "WomenP"
incl.cut <- 1
cov.cut <- 0
# truthTable
n.cut <- 1
incl.cut1 <- 1
incl.cut0 <- 1
complete <- TRUE
show.cases <- TRUE
sort.by <- "incl"
# minimize
explain <- 1
include <- ""
all.sol <- TRUE
rowdom <- FALSE
details <- TRUE
ttab.ne <- krook.ne
# positive case
neg.out <- FALSE
# superSubset
krook.susu <- superSubset(ttab.ne, outcome=outcome,
   incl.cut=incl.cut, cov.cut=cov.cut)
```

Para la minimización booleana mediante el paquete QCA de R y la función `minimize()`, primero creamos la tabla de valores de verdad mediante `truthTable()`. Esta función en última instancia sólo prepara los datos para el algoritmo y no cambia nuestra tabla de valores de verdad existente.

```
# truth table
# rows = unique configurations !!NOT cases
krook.TT <- truthTable(data=ttab.ne, outcome=outcome,
   neg.out=neg.out, n.cut=n.cut,
   incl.cut1=incl.cut1, incl.cut0=incl.cut0,
   complete=complete, show.cases=show.cases,
   sort.by="incl")
```

Es importante señalar que aquí las filas no representan casos, sino configuraciones condicionales únicas que pueden incluir varios casos. Ahora pasamos a la minimización booleana propiamente dicha. Si sólo nos interesa el resultado final sin especificar los parámetros de la función, la llamada es muy corta:

```
minimize(krook.TT)
```

Si quisiéramos controlar la función minimize() con más precisión y procesar el resultado aún más, el resultado es una llamada más larga:

```
# logic minimization
# do not use eqmcc anymore, call minimize with same parameters
krook.mini <- minimize(input=krook.TT, outcome=outcome,
   neg.out=neg.out, n.cut=n.cut,
   incl.cut1=incl.cut1, incl.cut0=incl.cut0,
   explain=explain, include=include,
   all.sol=all.sol, rowdom=rowdom,
   details=details, show.cases=show.cases)
```

Tabla 12.17: *Estudio Krook (2010, datos brutos)*

Country	Electtoral system	Quotas	Women's status	Women's movement	Left party strength	% Women in National Parliament
	(PR)	(QU)	(WS)	(WM)	(L)	(WomenP*)
Sweden	PR	Yes	Soc Dem	Non	5	47.3
Finland	PR	No	Soc Dem	Non	5	42.0
Norway	PR	Yes	Soc Dem	Autonomous	14	37.9
Denmark	PR	No	Soc Dem	Autonomous	7	36.9
Netherlands	PR	Yes	Conservative	Autonomous	7	36.7
Spain	PR	Yes	Conservative	Autonomous	0	36.0
Belgium	PR	Yes	Conservative	Autonomous	13	34.7
Austria	PR	Yes	Conservative	Non	8	32.2
New Zealand	Mixed	No	Liberal	Autonomous	9	32.2
Iceland	PR	Yes	Soc Dem	Autonomous	9	31.7
Germany	Mixed	Yes	Conservative	Autonomous	7	31.6
Switzerland	PR	Yes	Conservative	Non	4	25.0
Australia	Majority	Yes	Liberal	Autonomous	0	24.7
Luxembourg	PR	No	Conservative	Non	7	23.3
Portugal	PR	Yes	Conservative	Non	0	21.3
Canada	Majority	No	Liberal	Autonomous	0	20.8
United Kingdom	Majority	Yes	Liberal	Autonomous	0	19.7
France	Majority	Yes	Conservative	Autonomous	0	18.5
Italy	Mixed	Yes	Conservative	Non	3	17.3
United States	Majority	No	Liberal	Autonomous	0	16.3
Greece	PR	Yes	Conservative	Non	0	13.0
Ireland	Majority	Yes	Liberal	Autonomous	2	13.3

*Kriterium

Tabla 12.18: *Estudio Krook (2010, tabla de valores de verdad)*

Country	PR	QU	WS	WM	L	WomenP
Sweden	1	1	1	0	0	1
Finland	1	0	1	0	0	1
Norway	1	1	1	1	1	1
Denmark	1	0	1	1	1	1
Netherlands	1	1	0	1	1	1
Spain	1	1	0	1	0	1
Belgium	1	1	0	1	1	1
Austria	1	1	0	0	1	1
New Zealand	0	0	0	1	1	1
Iceland	1	1	1	1	1	1
Germany	0	1	0	1	1	1
Switzerland	1	1	0	0	0	0
Australia	0	1	0	1	0	0
Luxembourg	1	0	0	0	1	0
Portugal	1	1	0	0	0	0
Canada	0	0	0	1	0	0
United Kingdom	0	1	0	1	0	0
France	0	1	0	1	0	0
Italy	0	1	0	0	0	0
United States	0	0	0	1	0	0
Greece	1	1	0	0	0	0
Ireland	0	1	0	1	0	0

Para simplificar las cosas, hemos escrito una pequeña función `print.pis()` que recibe el resultado de `minimize()` como entrada y produce una salida ordenada de los resultados del análisis implicante.

```
print.pis(krook.mini)
```

Si comparamos los implicantes esenciales con el artículo original de Krook (2010, p.896), llegamos a una solución idéntica. Los implicantes primarios son:

```
> # extraer implicantes primarios
> paste(attr(krook.mini$PIchart, "dimnames")[[1]],collapse=" + ")
[1] "~PR*~WS*WM*L + PR*QU*~WS*WM + PR*QU*~WS*L +
     PR*QU*WM*L + PR*WS*~WM*~L + PR*WS*WM*L + QU*~WS*WM*L"
```

y los implicantes primarios esenciales resultan en:

```
> # extraer los implicantes esenciales
> paste(krook.mini$essential, collapse=" + ")
[1] "~PR*~WS*WM*L + PR*QU*~WS*WM + PR*QU*~WS*L +
      PR*WS*~WM*~L + PR*WS*WM*L"
```

Por último, podemos intentar una salida gráfica, siempre y cuando no tengamos demasiadas condiciones.

```
venn(krook.mini, ilabels=TRUE, col="blue",
   zcolor="steelblue, yellow, green",
   ellipse=TRUE, borders=FALSE, box=FALSE)
```

Esto da lugar a un diagrama de Venn (véase la Fig. 12.4, más arriba para el caso positivo), que los visualiza muy bien para configuraciones de condiciones más sencillas. Los diagramas de Venn más complejos con muchas configuraciones se vuelven tan confusos que a n = 6 ó 7 conjuntos se llega al final de las posibilidades para representar un análisis implicante de forma gráficamente comprensible. Los diagramas de Venn individuales se pueden representar fácilmente en forma abstracta.

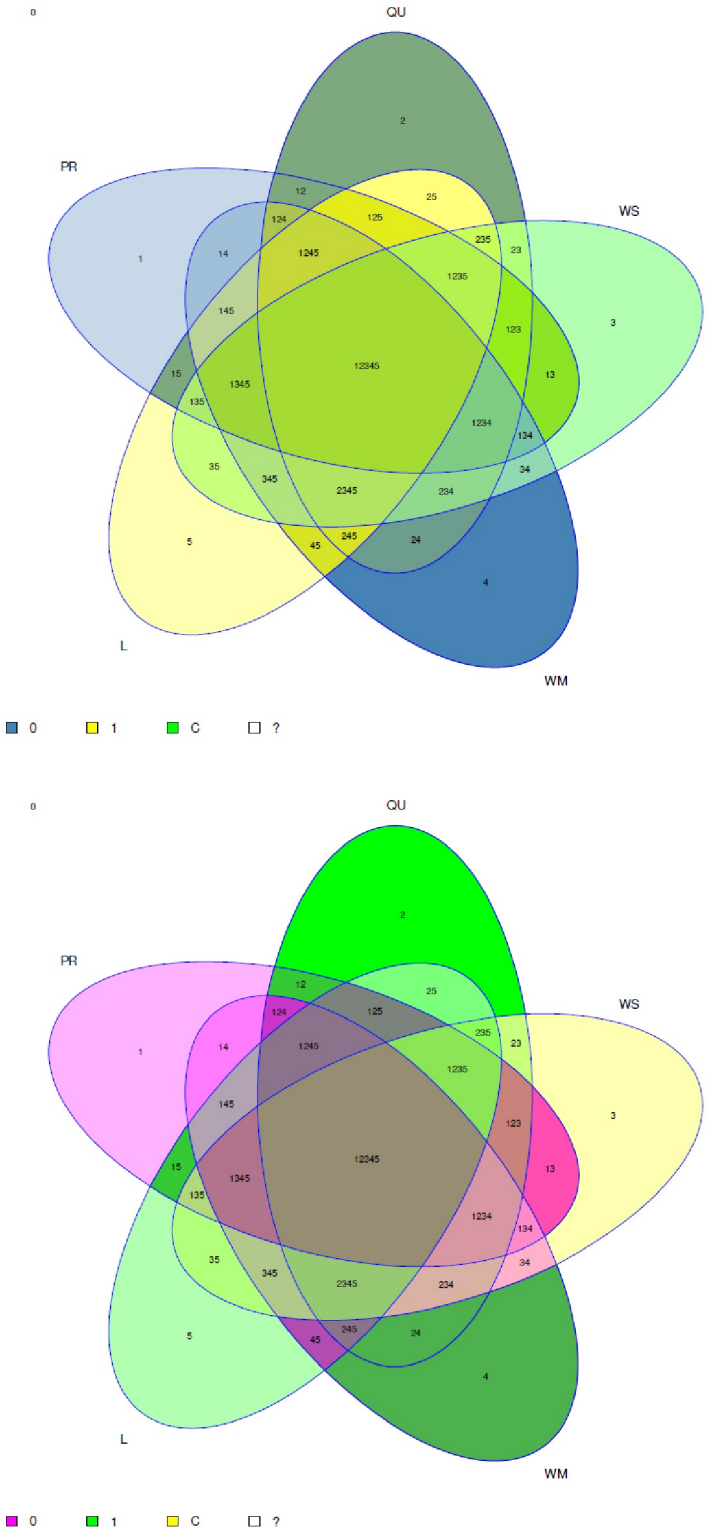


Figura 12.4: *Estudio Krook, 2010* (Diagramas Venn; Criterio: arriba VERDADERO abajo falso)

```
par(ask=TRUE)
for(i in 1:7) venn(i, ilabels = TRUE, zcolor = "style")
```

Ahora podemos empezar a interpretar las implicantes primarias (esenciales) y lo que cada una de estas configuraciones significa para el criterio de *representatividad de las mujeres en los parlamentos* (Krook, 2010, p.896). Antes de hacer esto, necesitamos evaluar el resultado negativo del criterio (Krook, 2010, p.897). Dejamos a los lectores como ejercicio los detalles del proceso completo de análisis R. Lo haremos breve (véase también la Fig. 12.4, más arriba).

```
# negative case
neg.out <- TRUE
# superSubset
krook.susu.NEG <- superSubset(ttab.ne, outcome=outcome,
   incl.cut=incl.cut, cov.cut=cov.cut)
# truth table
# rows = unique configurations !!NOT cases
krook.TT.NEG <- truthTable(data=ttab.ne, outcome=outcome,
   neg.out=neg.out, n.cut=n.cut,
   incl.cut1=incl.cut1, incl.cut0=incl.cut0,
   complete=complete, show.cases=show.cases,
   sort.by="incl")
# logic minimization
# do not use eqmcc anymore, call minimize with same parameters
krook.mini.NEG <- minimize(input=krook.TT.NEG, outcome=outcome,
   neg.out=neg.out, n.cut=n.cut,
   incl.cut1=incl.cut1, incl.cut0=incl.cut0,
   explain=explain, include=include,
   all.sol=all.sol, rowdom=rowdom,
   details=details, show.cases=show.cases)
```

Podemos imprimir y trazar todo con:

```
krook.mini.NEG R-Code
print.pis(krook.mini.NEG)
# plot
venn(krook.mini.NEG, ilabels=TRUE, col="blue",
   zcolor="steelblue, yellow, green",
   ellipse=TRUE, borders=FALSE, box=FALSE)
```

Echemos un vistazo más de cerca a los implicantes:

```
> # extract primary implicants
> paste(attr(krook.mini.NEG$PIchart,"dimnames")[[1]],collapse=" + ")
[1] "~PR*QU*~WS*~L + ~PR*~WS*WM*~L + QU*~WS*~WM*~L + PR*~QU*~WS*~WM*L"
> # extract essential implicants
> paste(krook.mini.NEG$essential, collapse=" + ")
[1] "~PR*~WS*WM*~L + QU*~WS*~WM*~L + PR*~QU*~WS*~WM*L"
```

Ahora sí que puede empezar el análisis comparativo y tener éxito. Antes comprobamos la coherencia, es decir, que no aparezca ninguna solución en ambos contextos:

```
> # check for consistency
> krook.pis <- attr(krook.mini$PIchart,"dimnames")[[1]]
> krook.pis.NEG <- attr(krook.mini.NEG$PIchart,"dimnames")[[1]]
> krook.pis %in% krook.pis.NEG
[1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE
> krook.pis.NEG %in% krook.pis
[1] FALSE FALSE FALSE FALSE
```

Esto es así porque las dos últimas llamadas sólo devuelven valores con falso lógico. Contrastemos ahora el caso positivo con el negativo: ¿bajo qué configuraciones condicionales las mujeres de los países occiden-

tales *entran o no entran* en el parlamento? Dejamos la interpretación del contenido a los lectores y remitimos una vez más al artículo de Krook para poder sentar las bases teóricas al respecto.

```
> # prepare positive and negative case for discussion
> krook.discuss <- primtab.desc(srctab=krook.ne, primtab=krook.mini,
   + outcome=outcome, norownames=FALSE)
negative outcome: FALSE
                PR QU WS WM L  [ WomenP ]
~PR*~WS*WM*L    pr -- ws WM L  WOMENP
PR*QU*~WS*WM    PR QU ws WM -- WOMENP
PR*QU*~WS*L     PR QU ws -- L  WOMENP
PR*QU*WM*L      PR QU -- WM L  WOMENP
PR*WS*~WM*~L    PR -- WS wm l  WOMENP
PR*WS*WM*L      PR -- WS WM L  WOMENP
QU*~WS*WM*L     -- QU ws WM L  WOMENP
> krook.discuss.NEG <- primtab.desc(srctab=krook.ne,
   + primtab=krook.mini.NEG,
   + outcome=outcome,
   + norownames=FALSE)
negative outcome: TRUE
                  PR QU WS WM L [ WomenP ]
~PR*QU*~WS*~L     pr QU ws -- l womenp
~PR*~WS*WM*~L     pr -- ws WM l womenp
QU*~WS*~WM*~L     -- QU ws wm l womenp
PR*~QU*~WS*~WM*L PR qu ws wm L womenp
```

En el curso del reanálisis de los datos de Krook, se observa que R produce *una solución más* para el caso del criterio negativo *Mujeres en el Parlamento* de lo que informa Krook (2010, p.897). Krook enumera el resultado = QU*ws*wm*l + pr*ws*WM*l + PR*qu*ws*wm*L. La configuración condicional adicional basada en nuestro análisis es Mujeres P = pr*QU*ws*l. Merecería la pena examinar con más detalle si Krook pasó por alto esta solución, su software Tosmana (Cronqvist, 2019) no la emitió – lo cual es poco probable – o hay otras razones pero no informadas. La solución parece pertinente en la medida en que el sistema político y la posición de las mujeres no desempeñan un papel para este resultado relativo a la ausencia de mujeres en los parlamentos. Se trata más bien de la presencia de cuotas de género y de la ausencia de partidos de orientación liberal -izquierda. En cuanto al contenido, habría que examinar más detenidamente las circunstancias para saber si las cuotas de género parecen desempeñar un papel inhibidor en la medida en que los partidos de orientación liberal-izquierda están ausentes del panorama partidista, y por qué.

Tarea 12.1: Mujeres en los parlamentos – Tarea complementaria

Otra tarea para los lectores interesados sería utilizar el álgebra de Boole para evaluar los datos publicados por Krook para el África subsahariana. Lamentablemente, el autor ha utilizado diferentes configuraciones de condiciones para estos países, de modo que una comparación directa de los países occidentales y africanos resulta imposible o va acompañada de enormes sacrificios de precisión. Los análisis para los estados africanos destacados se deberían implementar en R para ambos casos – criterio positivo y negativo – y comparar con los resultados de Krook. Se pueden encontrar las tablas necesarias en el artículo original de Krook o en nuestros scripts de R.

12.11.2 Vivir y morir en el Titanic — Parte II

El análisis del Titanic ya se ha realizado en detalle (véase el capítulo 5.5.4), por lo que pasamos directamente al análisis de las implicaciones. Los datos requerían una conversión a valores de verdad y hemos incorporado algunos detalles sobre los valores que faltan. Los detalles son evidentes en los scripts de R, implican código

R simple, y no entramos en ellos más adelante. Si lo desea, puede cambiar los datos usted mismo y reproducir los análisis. Dependiendo de los cambios, se verá que posiblemente los resultados también difieren. La tabla 12.19 ofrece una visión general de las variables utilizadas y las modificaciones que aplicamos a las variables, como el tratamiento de los NA = datos que faltan. Todas las variables son de tipo lógico y se han convertido en una matriz [0, 1] donde VERDADERO = 1 y FALSO = 0. Hay 8 variables con 2176 filas (personas).

Tabla 12.19: *Estudio Titanic (análisis de implicantes, variables)*

Variable	Variable en R	Valor en R*	Notas
Tripulación	`crew.c.L`	Miembro de la trip.	
Primera clase	`firstclass.1st.L`	Primera clase	
Clase más baja	`lowerclass.l.L`	Clase inferior	Se utilizó para el análisis, pero no para el diagrama de Venn, de lo contrario el número de categorías supera el tamaño representable de venn(), véase la categoría "primera clase".
Viajar sólo	`travelalone.a.L`	viajando sólo	
Sexo	`sex.f.L`	femenino	
Sobrevivir	`survived.s.L`	sobrevivido	
Entrada no en Southampton	`notjoinedinS.nS.L`	Entrada no en Southampton	
maternidad posible	`possible.mother.pm.L`	Posible madre	Los NA se establecieron como 0 (= sin madre).
Niño posible	`possible.child.c.L`	Posible niño	Los NA se establecieron como 0 (= ningún hijo).
Viajeros en grupo	`travel.hugegroup.4min.L`	viajando en grupo mín. 4 personas	No se utilizó para el análisis, véase la categoría "viajar solo").

* = TRUE

El resto del procedimiento sigue estrictamente el procedimiento del estudio de Krook (2010) ya presentado. No imprimimos los pasos individuales. Sin embargo, tras unos pasos, cabe señalar que hay claramente más soluciones para sobrevivir que para morir. Esto varía en función del conjunto de categorías incluidas (`ptIV_qual_Boole_case_Titanic_death-and-dying.r`).

```
# resultado y predictores en filas
> t(aquadstyle.rev(t.discuss))
                    1     2     3     4     5     6     7     8
tripulación.c.L     FALSE FALSE FALSE FALSE FALSE FALSE NA    FALSE
firstclass.1st.L    NA    NA    NA    NA    FALSE TRUE  FALSE FALSE
lowerclass.l.L      FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE
travelalone.a.L     FALSE FALSE FALSE NA    NA    TRUE  TRUE  TRUE
sex.f.L             NA    TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
notjoinedinS.nS.L   NA TRUE TRUE FALSE TRUE FALSE TRUE TRUE
possible.mother.pm.L FALSE FALSE NA FALSE FALSE FALSE FALSE FALSE
possible.child.c.L  TRUE NA FALSE TRUE FALSE NA FALSE TRUE
```

```
[ survived.s.L ]     TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
> t(aquadstyle.rev(t.discuss.NEG))
                     1     2     3
crew.c.L             FALSE FALSE FALSE
firstclass.1st.L     FALSE FALSE FALSE
lowerclass.l.L       FALSE NA    FALSE
travelalone.a.L      TRUE  TRUE  FALSE
sex.f.L              FALSE FALSE FALSE
notjoinedinS.nS.L    NA    TRUE  TRUE
possible.mother.pm.L FALSE FALSE FALSE
possible.child.c.L   TRUE  TRUE  FALSE
[ survived.s.L ]     TRUE  TRUE  TRUE
```

y completo como tabla 1/0:

```
> t.full.tab <- cbind( t(aquadstyle.rev(t.discuss)+0),
              + t(aquadstyle.rev(t.discuss.NEG)+0) )
> t.full.tab
                     1  2  3  4  5  6  7  8  1  2  3
crew.c.L             0  0  0  0  0  0  NA 0  0  0  0
firstclass.1st.L     NA NA NA NA 0  1  0  0  0  0  0
lowerclass.l.L       0  0  0  0  0  0  0  1  0  NA 0
travelalone.a.L      0  0  0  NA NA 1  1  1  1  1  0
sex.f.L              NA 1  1  1  1  1  1  1  0  0  0
notjoinedinS.nS.L    NA 1  1  0  1  0  1  1  NA 1  1
possible.mother.pm.L 0  0  NA 0  0  0  0  0  0  0  0
possible.child.c.L   1  NA 0  1  0  NA 0  1  1  1  0
[ survived.s.L ]     1  1  1  1  1  1  1  1  0  0  0
```

El resultado se muestra en la Tabla 12.20. La primera fila contiene el criterio, supervivencia SOBREVIVIDO. Los tipos P1 a P8 representan la salida positiva supervivencia y N1 a N3 la salida negativa muerte. Los tipos deben leerse en columnas, es decir, como la presencia o ausencia de la categoría respectiva, de modo que el criterio de la primera línea recibe su valor como una combinación lógica de las categorías. Estas combinaciones lógicas no se deben leer estadísticamente, sino como combinaciones integradas que conducen a un resultado lógico-causal. Esto significa que la presencia o ausencia de una categoría puede incluso conducir a ambos resultados – supervivencia o muerte – pero en una combinación diferente junto con otros factores.

Tabla 12.20: *Estudio Titanic (Análisis de implicantes)*

Categoría	P1	P2	P3	P4	P5	P6	P7	P8	N1	N2	N3
SOBREVIVIDO	1	1	1	1	1	1	1	1	0	0	0
crew.c.L	0	0	0	0	0	0		0	0	0	0
firstclass.1st.L					0	1	0	0	0	0	0
lowerclass.l.L	0	0	0	0	0	0	0	1	0		0
travelalone.a.L	0	0	0			1	1	1	1	1	0
sex.f.L		0	0	0	0	0	0	0	0	0	0
notjoinedinS.nS.L		1	1	0	1	0	1	1		0	0
possible.mother.pm.L	0	0		0	0	0	0	0	0	0	0
possible.child.c.L	1		0	1	0		0	1	1	1	0

Parte V

Síntesis

978

Capítulo 13
Combinar los Métodos

»Ich bin der Geist der stets verneint! / Und das mit Recht; denn alles was entsteht / Ist werth daß es zu Grunde geht; / Drum besser wär's daß nichts entstünde. / So ist denn alles was ihr Sünde, / Zerstörung, kurz das Böse nennt, / Mein eigentliches Element.«
¡Soy el espíritu que siempre niega! / Y eso con razón; pues todo lo que nace / Es digno de perecer; / Por lo tanto sería mejor que nada llegara a existir / Así pues, todo lo que llamais pecado, / Destrucción, en suma, mal, / Es mi real elemento.

de: Faust (Goethe, 1808, V. 1338–1344)
MEFISTÓFELES

13.1 En el país de la leche y la miel

El uso de métodos cuantitativos (CUAN) y cualitativos (CUAL) en la práctica de la investigación en las ciencias sociales se describe desde perspectivas epistemológicas muy diferentes. En la literatura encontramos autores que defienden una posición de contradicción irreconciliable entre CUAN y CUAL (por ejemplo, Smith & Heshusius,1986), en base a la cual los investigadores cuantitativos y cualitativos, en caso de discrepar, ni siquiera están en condiciones de discutir sus métodos entre sí y además con éxito. Otros afirman que los enfoques CUAL y CUAN se complementan de forma natural en el día a día (por ejemplo, Gürtler & Huber, 2006), lo que se refleja a nivel lingüístico en el hecho de que los estudios CUAN parten necesariamente de hipótesis significativas, formuladas esencialmente de forma cualitativa, relativas a los aspectos comunes y las diferencias, incluida la información sobre la dirección y el efecto, e intentan interpretar los resultados de una forma cualitativa igualmente significativa. En vista de la necesaria labor de traducción en el transcurso de un proyecto de investigación, este punto de vista parece natural (Gigerenzer, 1981; Gürtler, 2005). De hecho, en todo análisis cuantitativo entran siempre supuestos sobre la calidad particular de los hechos cuantificados. Se trata de suposiciones que nos permiten registrar los acontecimientos de la forma en que creemos que ocurrieron. A la inversa, la interpretación de textos incluye una gran variedad de enunciados cuantificadores (p. ej. con frecuencia, rara vez, siempre, nunca, ...) o comparaciones (por ejemplo, más, menos, más a menudo, menos a menudo, ...) relacionadas con las características cualitativas (por ejemplo, codificación, sistemas de categorías, hipótesis de secuencia, etc.). Estos enunciados describen la relación cualitativa con otras unidades de significado en el mismo texto o con otros textos o con las situaciones en las que se produjeron los textos, añadiendo un nivel de comparación sin el cual la investigación cualitativa perdería una dimensión importante. Así pues, la investigación CUAN no puede prescindir del apoyo CUAL, y viceversa. En las dos últimas décadas, por lo tanto, ha prevalecido cada vez más la postura de que es importante combinar sistemáticamente CUAN y CUAL (por ejemplo, Tashakkorie & Teddlie, 1998, 2003a) para utilizar de forma óptima las contribuciones específicas de los dos enfoques para responder a una pregunta de investigación.

Dado que la relación entre CUAN y CUAL puede acabar en una disputa escolar o en una sensación de total arbitrariedad ante la variedad de combinaciones disponibles, las voces críticas no son realmente sorprendentes. Por otra parte, la situación ha cambiado significativamente en las dos últimas décadas y los métodos mixtos gozan ahora de gran aceptación en la corriente dominante. Absurdamente, esto conduce a situaciones casi paradójicas: en el pasado, por ejemplo, sin duda habríamos pedido que se prestara más atención al tema de los métodos mixtos para responder a una pregunta de investigación de forma más exhaustiva y sobre una base más amplia. Hoy, sin embargo, se parece considerar o incluso esperar como buen estilo – especialmente en los trabajos de cualificación – utilizar métodos mixtos. Poner el método por encima del contenido y la intención de la investigación no es acertado, a menos que el trabajo sea puramente metodológico. Una expectativa irreflexiva qua convención es tan poco inteligente como prescindir de ella por completo; y esto a su vez es similar a la discusión sobre la significación que se trata ampliamente aquí en el libro (véanse, entre otros, los capítulos 4.3.9.1 y 4.3.8) en la estadística clásica y la creciente demanda de factores de Bayes (véase el capítulo 6.8.1.7 para una discusión resumida) en la estadística de Bayes. *Más métodos no significa más calidad*. El uso de métodos de fuerza bruta no garantiza la calidad en absoluto: si ajustamos un telescopio desenfocado, veremos el fenómeno borroso desde todos los puntos desde los que lo observemos. Entonces, es evidente que preferiríamos un telescopio bien ajustado desde un punto de vista único pero adecuado a todos los demás puntos de observación desenfocados. Por otro lado, si podemos desplazar el telescopio bien ajustado a puntos de observación que son estratégicamente importantes para nosotros, naturalmente preferiríamos la medición nítida múltiple a una única. Pero entonces deberíamos saber exactamente en qué puntos de observación nos encontramos y qué significan en términos de contenido para poder establecer las conexiones. Y en eso consiste la conexión entre QUAL y QUAN: *anclar el contenido en el objeto*. Si esto es transparente y reproducible o comprensible, la ubicación categórica de la metodología según CUAN, CUAL, lógica u otra cosa nos parece de importancia secundaria y terciaria. En este sentido:

Recordatorio 13.1: Influencia de la pregunta de investigación

La pregunta de investigación determina la elección de los métodos.

Por lo tanto, si se puede responder una pregunta de investigación ya de forma exhaustiva mediante el uso exclusivo de estadísticas, no hay razón para combinar métodos. Lo mismo ocurre si un enfoque cualitativo ya es suficiente. Los métodos deben combinarse si el uso combinado añade valor a la respuesta a la pregunta de investigación, lo que no ocurriría sin esta combinación.

El peligro de los *métodos mixtos* es que, en determinadas condiciones, pueden dar lugar a un auténtico batiburrillo, es decir, a una papilla poco clara en la que todo se mezcla y el trigo y la paja ya no pueden separarse claramente entre sí. En este sentido, es preferible hablar de *combinación de métodos*, en primer lugar porque así se evita la idea de mezcla, ya que no se trata de eso. No estamos preparando un batido en el que la entropía sea máxima (véase el capítulo 6.14.2). El énfasis en la *combinación* deja claro que lo que estructuralmente deben y pueden combinarse entre sí si encajan estructuralmente y en términos de contenido, y si prometen una ganancia de conocimiento. Además, la idea de CUAN y CUAL no está representada en el nombre, porque se puede combinar mucho más que CUAN y CUAL. En la tabla 13.1 enumera las combinaciones posibles sólo para el nivel de los métodos de análisis de datos y que a lo largo de la subdivisión elegida aquí en el libro. Otras subdivisiones son en principio otras subdivisiones y la tabla 13.1 podría ampliarse. Si ahora nos preguntamos por qué Neyman-Pearson puede combinarse con Fisher, la respuesta es - porque se puede, en principio cualquier conjunto de datos puede, en principio, examinarse con cualquier método. Esto demuestra que la combinación de métodos nunca es el problema, porque no dice nada sobre si tal comparación o combinación tiene sentido. Lo que siempre se combina, es decir, qué datos de qué periodo de encuesta conmuestra, permite un campo de infinitas posibilidades. Por ejemplo, se podría combinar la estadística clásica frente a Bayes para examinar si las respectivas conclusiones prácticas son congruentes. De este modo, se podría saber cuál es la parte del método de análisis de datos y cuál es la parte real de los datos. Dentro de la estadística clásica, por ejemplo, se podrían examinar los mismos datos con diferentes métodos de análisis (por ejemplo, análisis de conglomerados) para comparar los resultados a nivelde conclusiones prácticas y teóricas, por un lado para explorar los algoritmos utilizados en el material de datos y, por otro, para hacer de la tolerancia o estabilidad de los propios datos el tema. La estadística de

Bayes podría combinarse con el análisis de secuencias para evaluar las intervenciones terapéuticas, como hizo Studer (1998), por un lado, y por otro, para investigar la calidad de los datos en la línea de un análisis en profundidad. La estadística de Bayes podría combinarse con el análisis secuencial para evaluar las intervenciones terapéuticas, por un lado, y, por otro, para examinar la calidad de los procesos de cambio examinados a lo largo de una profundización. El análisis implícito, por otro lado, sería una forma de contrastar estos procesos de cambio con modelos estadísticos complejos orientados a los análisis de impacto. Una vez más, la cuestión es: ¿congruencia o no de los resultados? Esta es sólo una modesta del espectro de posibilidades. Desde nuestro punto de vista, se pueden combinar siempre que se realicen siguiendo estrictamente una pregunta de investigación razonable y científica y se atengan a las normas científicas pertinentes. Que los criterios de calidad difieran más o menos, ya que se utilizan términos diferentes es un reto, pero en ningún caso un obstáculo.

Tabla 13.1: *Posiblidades de combinar los métodos del análisis de datos*

		CUAN			CUAL		Lógica
		N-P	Fisher	Bayes	KodierP.	SeqA.	BM
CUAN	N-P	×					
	Fisher	×	×				
	Bayes	×	×	×			
CUAL	KodierP.	×	×	×	×		
	SeqA.	×	×	×	×	×	
Lógica	BM	×	×	×	×	×	×

Abreviaciones:

N-P = Neyman-Pearson,
KodierP. = Paradigma de codificación,
SeqA. = Análisis de secuencia (Hermeníutica Obj.),
BM = Minimización booleana (An. de implicantes)

Es importante que se puedan combinar los métodos siempre entre sí (por ejemplo, mediante diferenciación o cambios en el diseño, otros datos, etc.). Además, no sólo se pueden combinar los métodos de análisis de datos entre sí, sino todos los elementos del proceso de investigación (sobre la triangulación Flick, 2000): Teorías, hipótesis y conjeturas, muestras, instrumentos de recogida de datos y métodos de análisis, y resultados (interpretaciones). En principio, no hay límites, lo que hace que la justificabilidad sustantiva parezca aún más relevante. Si uno se orienta por el todo vale de Feyerabend (1976, véase el capítulo 3), existe incluso la posibilidad de vincular fuentes de conocimiento ajenas a lo que generalmente se acepta como científico. Todo el proceso de combinar métodos se parezca tanto a un metaanálisis como a un gran supermercado con demasiadas posibilidades. La referencia directa al metaanálisis es obvia porque, dado el uso de múltiples métodos, se dispone de múltiples análisis y resultados incluso del mismo tema (es decir, desde diferentes perspectivas) que es necesario integrar. Esto es precisamente lo que hace el metaanálisis: integrar los resultados de distintos estudios para determinar los efectos globales. La cientificidad resulta cuando se seleccionan de esta tierra de leche y miel aquellas pocas combinaciones que realmente se ajustan a un problema de investigación. No hay que limitarse en esto, porque la ciencia vive de la creatividad y de nuevas ideas para probarlo. La adhesión estricta a las convenciones salvaguarda el statu quo, pero noningún territorio nuevo. Este punto de vista puede resumirse así:

Recordatorio 13.2: Combinación de métodos y metaanálisis.

Entendemos las combinaciones de métodos como una forma ampliada del metaanálisis, en la que uno o más objetos de investigación se examinan desde diferentes perspectivas. El objetivo es reunir los (múltiples) resultados – incluso de un único estudio – en una estructura unificada que esté abierta a la comparación, con el fin de integrar plenamente estos resultados. Esto debería permitir responder a una pregunta de investigación de forma adecuada a cada caso haciendo uso de toda la información disponible.

Las combinaciones de métodos se refieren al uso de instrumentos de investigación y/o métodos (de análisis de datos) específicos del contexto y, como tales, se tiene que *derivarlas directamente de la pregunta de investigación y fácticamente justificarlas en ella*. Esto, a su vez, no debería imponer ninguna restricción adicional, sino simplemente subrayar la *importancia de la justificación*. Aquí no hay lugar para convenciones, expectativas o sensibilidades de una determinada *comunidad social*, porque eso significaría negar o devaluar masivamente la relevancia de la pregunta de investigación en favor de las expectativas sociales de otros y en perjuicio de la investigación como tal. Este sería un punto de partida bastante modesto para el progreso científico, lo que no significa que exactamente esto no ocurra – con frecuencia – en la práctica. Sin duda, todo trabajo de investigación y, en particular, los trabajos de cualificación tienen limitaciones más o menos claras en cuanto a organización y recursos. La exigencia de una justificación sustantiva estricta a la hora de combinar métodos no es justificación de la combinación de métodos no es su objetivo. Simplemente hace hincapié en el trabajo de justificación que siempre debe realizarse. Esto no significa, sin embargo, que los criterios de calidad del trabajo científico y no excluye la posibilidad de realizar cambios de gran calado. Los cambios en el ámbito de los criterios de calidad requieren naturalmente requieren una justificación realmente buena.

13.2 Sobre la complementariedad de los métodos cuantitativos y cualitativos

Un ejemplo de un estudio exploratorio de los procesos de *aprendizaje en una clase de geometría* de 10º curso de secundaria (Tinto, 1986) ilustra el problema de la aplicación unilateral de los métodos. En los protocolos de observación, el *volunteering* (participación voluntaria) destaca entre las categorías para registrar el comportamiento de los alumnos. La autora utiliza este término para referirse a la *participación espontánea* de los alumnos en el aula. Los datos se reducen, es decir, un gran número de sentencias se sustituye con la ayuda de reglas de mapeo por un conjunto mucho más pequeño de sentencias, posiblemente de símbolos únicos. En el caso de las clases de geometría, todas las secuencias de protocolo que, por ejemplo, incluyan los siguientes acontecimientos, podrían reducirse a la única categoría de *participación espontánea*:

1. El profesor hace una pregunta a toda la clase,
2. ... describe un problema,
3. ... se refiere a un paso de la prueba ... ... y mira a la clase,
4. ... formula una pregunta complementaria, a la que un alumno
   a) responde,
   b) formula una explicación,
   c) sugiere una solución,
   d) pide detalles
5. [...]

Sólo según esta lista, son concebibles 24 variantes de interacción para cada alumno; todas estas secuencias posibles se reducen al dato *participación espontánea/alumno XYZ* y pierden así su nivel original de interpretación en términos de contenido. Un análisis cuantitativo de las categorías reveló entonces grandes diferencias en la frecuencia de la participación espontánea en el artículo de Tinto (ibíd.).

Sin embargo, el análisis no puede agotarse en esta observación, ya que una gran variabilidad sólo permite unas pocas derivaciones estrechamente definidas, que sin embargo son relevantes. *Frecuente* e *infrecuente* por sí solos no son suficientes aquí, para comprender el comportamiento de los alumnos en el contexto de su formación. ¿Qué debemos hacer con los valores de tendencia central y dispersión de la participación espontánea tomados por sí solos?

Estos valores sólo adquieren sentido cuando podemos relacionarlos con otras características de la lección o consecuencias para los alumnos. Una tabulación cruzada de las cantidades de rendimiento y participación espontánea podría sugerir que esta forma de participación no tiene nada que ver con el éxito de la enseñanza. Por ejemplo, de los dos mejores alumnos de la clase, Sandra casi siempre hablaba de la forma categorizada, mientras que este comportamiento casi nunca se observaba en Marcos (los nombres son anónimos). Ya una variabilidad superficialmente máxima se desarrolla bajo la misma categoría de contribución espontánea. Sin embargo, Tinto (ibíd.) no sólo dispone de protocolos de observación, sino también de entrevistas con alumnos y con el profesor de matemáticas. De los textos anteriores se desprende que la participación espontánea puede tener significados muy diferentes desde la perspectiva de los alumnos. Mientras que Sandra participa por voluntad propia *para garantizar el progreso de la clase* y evitar el cansino sarcasmo del profesor ante la falta de cooperación (por ejemplo, ¡Ahora todos asientan con la cabeza!), Mark sólo participa espontáneamente *cuando espera la confirmación de que su punto de vista, solución, etc. son realmente correctos*. ¿Pueden leerse estas definiciones cualitativamente diferentes a partir de los datos cuantitativos? ¿Deben combinarse ambas en la misma categoría? Definitivamente no, y esto no se debe a las estadísticas, sino a que estos análisis no son adecuados para de los datos que aquí se presentan. Pero al comprender que la visión puramente estadística de las cosas no es suficiente, uno hace exactamente lo que sugiere Tukey (1980, véase el capítulo 5): Se explora los datos, juega con ellos y descubre que el nivel numérico es insuficiente. De este modo, uno interpreta cualitativamente las cantidades calculadas y las rechaza por insuficientes. Esto no significa que se descarte por completo este punto de vista, sino que simplemente se va más allá y se recurre a las entrevistas, por ejemplo, que entonces -¡oh maravilla! - muestran muy claramente por qué Sandra y Mark se comportan de manera muy diferente, aunque a nivel superficial su comportamiento pertenezca a la misma categoría contable. Una vez alcanzado un nivel adecuado de categorías o sistemas de clasificación con los análisis cualitativos posteriores, tiene sentido volver a contar y comparar frecuencias o estimar probabilidades y relacionarlas entre sí de forma significativa.

Este esquema de análisis debe dejar suficientemente claro que no se pueden ni deben analizar los textos ni cuantitativamente ni cualitativamente. El recuento de las frecuencias de una categoría presupone un trabajo cualitativo previo de definición precisa de dicha categoría. Las cantidades resultantes deben interpretarse recurriendo a su calidad. Podríamos preguntarnos: ¿cómo queremos distinguir unos acontecimientos de otros? ¿Qué progreso en el conocimiento nos aporta la determinación de la frecuencia de determinados sucesos si no queremos atribuirles determinadas cualidades? Los supuestos sobre el carácter particular de los sucesos cuantificados siempre fluyen en todo análisis cuantitativo.

A la inversa, la denominación de rasgos cualitativos del texto sin información cuantitativa (por ejemplo, frecuentemente, raramente, siempre, nunca) es igual de difícil. La única excepción es el análisis de secuencias (véase el capítulo 11.2 o el capítulo 11.9 sobre metodología práctica), que en realidad funciona sin cantidades ni términos cuantitativos. Aunque sólo se tratara de mostrar cómo piensan los alumnos sobre la cooperación espontánea, seguiríamos sin poder prescindir de la cuantificación. ¿Quién querría leer y comparar una serie de unas 50 reconstrucciones de significados subjetivos, y qué ganaría el lector con esta lectura, en algunos aspectos bastante redundante y posiblemente muy aburrida? Más le valdría al lector hacer él mismo el análisis. Debería haber cierta eficacia y rigor. Así pues, un lector que siguiera interesado tras la lectura completa de dichas 50 reconstrucciones haría lo que el investigador debería haber hecho de inmediato para complementar su análisis cualitativo: A saber, producir una visión de conjunto ordenando según ciertos patrones de significado para abarcar y explicar bien todo el campo de investigación a partir de unos pocos rasgos estructurales. Sin embargo, esta simple clasificación no es más que una primera forma de cuantificación. Que se hable de configuraciones de sentido X, Y o Z (en formulaciones más o menos detalladas) o que simplemente se asignen los números 1, 2 y 3 para etiquetar, no supone diferencia alguna.

En la clasificación o tipologización, se asignan los resultados cualitativos a una escala nominal, la cuantificación o nivel de escala más simple. En este caso, la reducción cuantitativa complementa y enriquece el análisis cualitativo al transmitir claridad, por lo que adquiere un lugar muy valioso en la estructura general. Para garantizar la calidad de estas estrategias de combinación de métodos, parece importante, al utilizar procedimientos cuantitativos, abrir y revelar cualitativamente las dimensiones de significado de los datos y divulgarlas. Cuando se utilizan métodos cualitativos, es importante sistematizar y documentar los procesos

de interpretación y, además, ordenar los hallazgos de forma cuantitativa. Si se cumple este requisito, entonces es posible y necesario utilizar los hallazgos y procedimientos de la otra orientación metodológica en cada una de las dos orientaciones aparentemente incompatibles; y entonces ya no puede encontrarse una supuesta contradicción entre CUAN, CUAL y otros enfoques. Los requisitos para ello son la transparencia y la regularidad, la metodología real en la aplicación práctica de los enfoques metodológicos en la investigación. En este sentido, no existe investigación dura o blanda, sino, independientemente de la calidad del anclaje teórico de un estudio, sólo rigor metodológico, selección y aplicación casuística de los métodos o, por desgracia, dejadez, en ambos enfoques. Un grupo de trabajo de la Asociación Americana de Psicología (APA, 2006) señaló en general la necesidad y la importancia de basar la acción práctica en hallazgos científicos obtenidos mediante métodos tanto cuantitativos como cualitativos. Iniciar una disputa escolástica sobre la superioridad de uno u otro arsenal de métodos sólo puede calificarse de burdas tonterías. Los métodos sólo son tan buenos como la calidad de su aplicación y su justificabilidad sustantiva ante una pregunta de investigación. En cualquier caso, los métodos sólo deben evaluarse si alguien los ha practicado realmente (todos), al menos aquellos sobre los que se hacen afirmaciones evaluativas. Éste es el problema de muchos filósofos de la ciencia como Popper, que desgraciadamente no llevaron a cabo una amplia investigación empírica en un contexto definido. Sólo los experimentos de pensamiento no son suficientes, y así ciertos pensamientos nunca pueden ser adecuadamente pensados y probados en la práctica – uno piensa en el argumento antibayesiano de Popper (1943), que ha sido refutado por Jaynes (2003).

Por lo tanto, los métodos cualitativos y cuantitativos no se excluyen mutuamente, sino que tienen una *relación complementaria*. En primer lugar, es cierto que la elección del método depende de la pregunta a la que se pretenda responder. Si se quiere estimar cuántas personas votarán a un determinado partido o querrán comprar un determinado producto en el transcurso de los próximos seis meses, a uno le interesan como resultado los valores numéricos más fiables posibles. Si, por el contrario, quieren averiguar por qué la gente quiere comprar un determinado producto o por qué prefiere el producto de un competidor, preferirán evaluaciones diferenciadas desde el punto de vista subjetivo de los clientes y combinarlas con factores externos. Si se observan con detenimiento los dos estudios expuestos, en los que en el primer caso hay un valor porcentual con indicación de los límites de confianza, es decir, una indicación cuantitativa, y en el segundo caso una recopilación de cualidades atribuidas y factores de influencia, se advierte que en ambos estudios se producen actividades cualitativas y cuantitativas. Y esto no contradice que posiblemente el enfoque y el énfasis en cada caso tienda más hacia un arsenal de métodos u otro. Si la elección de los métodos surge de forma natural de la pregunta de investigación, no se pregunta por el tipo de método (es decir, CUAL, CUAN, ...), sino si se ajusta al contenido. Sin embargo, en el primer caso el papel de los métodos cualitativos y en el segundo el de los cuantitativos no suelen ser tratados en profundidad por los respectivos investigadores en sus artículos, de modo que surge la ilusión de que la investigación sólo se desarrolla dentro de un determinado paradigma. Esto raramente es así. El punto de partida de la investigación es un problema, en las ciencias sociales, normalmente un problema en un ámbito de la vida o un campo de acción social. La percepción más o menos difusa de este problema adquiere sus contornos únicos en la formulación de una o varias preguntas de investigación. En estas fases, las hipótesis no se comprueban cuantitativamente, sino que primero se generan las preguntas adecuadas y luego se las especifican. En la exploración posterior, en los estudios cuantitativos se deben establecer ahora las hipótesis y redactar un diseño de investigación, así como seleccionar los métodos con los que sea posible la comprobación (cuantitativa) de las hipótesis. En el caso de los estudios cualitativos, el diseño se centra en las posibilidades metodológicas para obtener más información sobre el área del problema para poder formular hipótesis científicas al respecto. Tanto si se utilizan tests, escalas de valoración, cuestionarios, entrevistas, diarios de revisión, grabaciones de vídeo y audio, imágenes, etc. – en cualquier caso, es necesario tener acceso al campo en el que se puede recopilar la información. Establecer contactos sobre el terreno, generar confianza y crear voluntad de participar en el estudio son otras actividades de investigación necesarias más allá de la visión polarizadora de los métodos cualitativos y cuantitativos.

Sólo queremos mencionar, pero no discutir más, el hecho de que, por supuesto, los instrumentos mismos que se utilizan en los estudios cuantitativos en la siguiente fase para explicar el problema no se desarrollan exclusivamente con métodos cuantitativos sino que ya tienen el elemento cualitativo como parte integrante. Especialmente cuando se construye un test altamente estandarizado (por ejemplo, un test de CI), alguien debe haber tenido la idea en algún momento y haber decidido que los ítems del test podrían captar lo que se supone que el test debe medir y hacerlo con un grado muy alto de precisión, lo que requiere una planificación cuidadosa y muchos ensayos. Y sobre todo en el caso de pruebas muy estandarizadas, como los tests de CI, con una gran fiabilidad y una validez que posiblemente permita incluso afirmaciones y recomendaciones individuales, suponemos que una gran parte del trabajo de desarrollo se ha dedicado

incluso a esta área cualitativa. E incluso con un enfoque de *fuerza bruta* como el MMPI (Minnesota Multiphasic Personality Inventory, Hathaway & McKinley, 1940), donde los ítems se generaron originalmente de forma *puramente empírica* sobre grupos categorizados clínicamente (operacionalizados como la probabilidad de respuesta a un ítem relacionado con un subgrupo categorizado clínicamente específico), una interpretación posterior requiere la parte cualitativa. Si no se desea interpretar, se pueden identificar las agrupaciones por separado unas de otras sobre una base probabilística, pero entonces uno no tiene ni idea de *por qué* esto es así o *qué significa*. En el caso del MMPI, esto puede ser útil para hacer distinciones al margen de consideraciones teóricas, pero el criterio de estas distinciones siguen siendo – originalmente – las clasificaciones clínicas subyacentes relacionadas con la muestra de calibración. Y estos *diagnósticos* se derivan a su vez de un proceso de construcción cualitativa, a saber, la clasificación de una persona en un trastorno, estructura de personalidad, etc. La característica de la supuesta *ausencia de teoría* no está realmente libre de teoría, sino que sólo lo están las derivaciones de la misma – es decir, las distinciones sobre la base de los ítems – y ello sólo mientras no se realicen derivaciones en cuanto al contenido que requieran interpretaciones – es decir, que en realidad no lo requieran en absoluto. Si a continuación se examina la literatura sobre cómo se debe utilizar el MMPI, las distinciones teóricas y, por tanto, cualitativas, relativas al significado de las escalas clínicas básicas, entran por la puerta de atrás. Así pues, en cuanto se trata de la aplicación de un instrumento de investigación y, por tanto, en general, de la interpretación de los resultados empíricos en un contexto definido, es completamente imposible trabajar sin el elemento cualitativo, a saber, la interpretación y la referencia. Cualquiera otra cosa sería muy sorprendente, porque los datos no se interpretan a sí mismos por sí mismos.

13.3 Modelos de combinación de métodos

13.3.1 CUAL y CUAN en diseños de conversión

Teddlie y Tashakkori (2006) han propuesto una taxonomía de combinaciones de métodos que incluye un modelo por el que se realizan análisis cualitativos y cuantitativos en el mismo conjunto de datos. Se trata de convertir datos cualitativos en cuantitativos y viceversa. Por lo tanto, los autores hablan de diseños de conversión y los describen como un modelo especial de combinación de métodos en el que los datos se cuantifican o cualifican (ibíd., p. 17). Tales combinaciones mediante la conversión de datos siempre se han llevado a cabo tanto en estudios cuantitativos como cualitativos, sin ninguna han surgido guerras de paradigmas (Gage, 1989; Bryman, 2008), como las de Smith y Heshusius (1986) con su tesis de que los investigadores cualitativos y cuantitativos ni siquiera pueden hablar entre sí debido a las incompatibilidades epistemológicas de los enfoques. Esta posición extrema asume implícitamente que en la investigación cuantitativa no hay necesidad de interpretar los datos y los resultados y que en la investigación cualitativa no hay cuantificación de ningún tipo. Del mismo modo, podríamos preguntarnos si los investigadores pertenecen siempre a un único paradigma de por vida o si están relacionados con un contexto específico, si los psicoanalistas nunca refuerzan y, por lo tanto, nunca utilizan técnicas de terapia conductual... Sin embargo, cualquier cantidad sólo aporta algo para responder a preguntas de la investigación sólo si los investigadores pueden interpretarlas en términos de presupuestos teóricos – y son precisamente tales interpretaciones las que son cualitativas por naturaleza. Por otra parte, cualquier informe sobre comparaciones cualitativas depende de que se realicen al menos algunas cuantificaciones mínimas se llevan a cabo, por ejemplo: Los participantes en la entrevista generalmente/ rara vez/ nunca creían que...

La opinión que se esgrimió repetidamente en la década de 1980, a saber, que los métodos cualitativos y cuantitativos eran incompatibles en el proceso de investigación, ignoraba por completo un desarrollo que se venía produciendo desde algunos años antes. Tukey (1977) publicó el trabajo seminal sobre el *análisis exploratorio de datos* (AED; acrónimo angloamericano EDA), que se ocupa de maximizar la comprensión de un conjunto de datos por revelar estructuras ocultas en los datos, extraer variables importantes, identificar casos atípicos y anomalías, probar conjeturas, desarrollar modelos parsimoniosos y determinar combinaciones óptimas de factores (National Institute of Standards and Technology, NIST, 2003). El NIST explica además que la verdadera percepción y 'sensación' de un conjunto de datos llega cuando el analista sondea y explora juiciosamente las diversas sutilezas de los datos (ibíd., párrafo 1.1.4, sin números de página). Para ello, es especialmente importante recurrir a nuestras propias capacidades humanas de reconocimiento de patrones y de comparación en el contexto de una serie de técnicas gráficas juiciosas aplicadas a los datos (ibíd., párr. 1.1.4).

El AED engloba estrategias estadísticas que permiten detectar regularidades o patrones específicos en los datos sin aplicar inmediatamente métodos confirmatorios complejos. En el contexto de la investigación cualitativa, parece especialmente interesante la contribución que el AED puede hacer al objetivo central de generar hipótesis. Alea, Jiménez, Muñoz, Torrelles y Viladomiu (2014, p.63) subrayan en su introducción que mediante AED será posible identificar medios de síntesis apropiados para resumir la información contenida en la muestra y formular hipótesis sobre la población. Del mismo modo, Inzunsa (2018) describe la utilidad del AED en un estudio empírico sobre razonamiento estadístico realizado por sus estudiantes universitarios. El análisis exploratorio de datos puede ser una herramienta útil para generar hipótesis, conjeturas y preguntas sobre los fenómenos de los que se recogieron los datos, y por eso se propone como un medio para desarrollar una comprensión global de los datos (ibíd., p.1264). Así pues, el AED promete apoyar los objetivos esenciales de la investigación cualitativa. En palabras de Flick (2009, p.12): El objetivo de su investigación no es tanto poner a prueba lo que ya se sabe (p. ej, teorías ya formuladas de antemano), sino descubrir y desarrollar lo nuevo y elaborar teorías empíricamente.

Los métodos de AED ofrecen diversas representaciones gráficas de los resultados de las operaciones cuantitativas, con las que se pueden representar patrones y estructuras más o menos ocultos en los datos. El adagio de que una imagen puede sustituir mil palabras se confirma aquí, porque los gráficos producidos por AED ofrecen literalmente una visión de la estructura de los datos. Los gráficos, elaborados con métodos cuantitativo-estadísticos, permiten a los investigadores identificar fácilmente correlaciones, similitudes y diferencias entre sus casos. Por otra parte, apoyan la comprensión de las conclusiones de sus informes de investigación. Un ejemplo muy sencillo: observamos qué estructuras surgen en los números del 1 al 30 si son todas vinculadas de modo aditivo, sustractivo, multiplicativo y por división. Algunos R-código muestra esto:

```
outer(1:30,1:30, +)
outer(1:30,1:30, -)
outer(1:30,1:30, *)
round(outer(1:30,1:30, /),2)
```

Si nos alejamos un poco de la pantalla y observamos las tablas numéricas un poco desenfocadas, podemos ver las *líneas rectas* en el caso de los enlaces aditivos y las *curvas* en el caso de los enlaces multiplicativos. Del mismo modo, podemos ver el *ascenso* de líneas y curvas en el caso de la suma y la multiplicación, y el *descenso* en el caso de la resta y la división. A partir de esta información, podríamos formular hipótesis sobre el efecto de la suma/resta, por un lado, y de la multiplicación/división, por otro, sin tener que saber mucho al respecto. Esta sencilla forma de ver los datos visualmente ya es posible en la escuela primaria y constituye una buena introducción al AED.

Así, es fácilmente posible calcular números primos no sólo en R con una sola línea – de nuevo, algo ineficiente y de memoria intensiva, pero ese no es principalmente el punto aquí. Empezamos la función para los números hasta el límite con la llamada `limit` sobre la base del algoritmo euclidiano y trabajando con la división entera:

```
> limit <- 100
> (res <- sapply(seq_along(2:limit),
+ function(i) ifelse(sum((i %% 2:(i-1)) == 0) == 0,i,NA)))[!is.na(res)]
[1] 3 5 7 11 13 17 19 23 29 31 37 41
[13] 43 47 53 59 61 67 71 73 79 83 89 97
```

Aquí, sin embargo, no se ve tanto en la salida numérica para combinar esta información con, digamos, información visual. Esto sería posible si los números primos se dispusieran en un sistema de coordenadas, como demostró de forma impresionante Sanderson (2019), y que entonces revelan una forma de espiral.

En el campo de la investigación cuantitativa, puede ser beneficioso complementar los enfoques clásicos con inferencias interpretativas que se basen en representaciones gráficas de AED. Por ejemplo, un simple gráfico puede responder a la pregunta de si existe o no una interacción entre los predictores de un modelo lineal, dificultando o no la interpretación de los predictores individuales (véase `?interaction.plot`). Sería mucho más complicado basar la misma conclusión únicamente en los coeficientes del modelo lineal. En general, las evaluaciones gráficas son de gran importancia en el caso de los modelos lineales (jerárquicos) y la evaluación de su bondad y ajuste a los datos. Los coeficientes por sí solos no bastan para ello, especialmente cuando se trata del tipo y la naturaleza de correlaciones que sólo un gráfico puede ilustrar (como el paquete R `HLMdiag`, Loy & Hofmann, 2013 o `ggfortify`, Yuan, Horikoshi & Li, 2016). Los

umbrales pueden definirse para identificar valores atípicos o desviaciones de las normas asumidas, o simplemente para comprobar el ajuste de los residuos a la distribución normal.

Por supuesto, en los estudios cualitativos, se necesita principalmente datos cuantitativos para poder aplicar métodos de AED y generar representaciones gráficas. Por lo tanto, a continuación examinaremos las formas de conversión de datos, especialmente la cuantificación de datos cualitativos. Se distinguen *tres formas de conversión de datos*.

13.3.2 Formas de conversión de datos

La *primera* forma de conversión de datos es tan habitual en la investigación en ciencias sociales que en realidad nadie discute la conversión de CUAL a CUAN o viceversa. La *segunda* forma se basa en la conversión de datos mayoritariamente cualitativos en datos de frecuencia. Por último, la *tercera* forma combina un análisis cualitativo completo y un posterior AED de los datos cualitativos convertidos en cantidades (frecuencias o probabilidades).

13.3.2.1 Conversión de datos cotidianos

Aunque la combinación de datos cualitativos y cuantitativos suele suscitar preocupación, esta forma de conversión es *tan habitual que nadie cuestiona* esta combinación de métodos (Teddlie & Tashakkori, 2006, p.19). En muchos estudios, esta conversión se utiliza para agrupar casos, personas, contextos o similares, y así combinar segmentos específicos del conjunto de datos entre sí. En la mayoría de las investigaciones en ciencias sociales, los datos socio-demográficos incluyen datos métricos, como la edad de los entrevistados/encuestados o la duración de su experiencia laboral. Normalmente, estos datos se agrupan y se convierten en valores nominales. Por ejemplo, se podrían comparar tres grupos de encuestados con experiencia laboral baja/media/alta. Por otra parte, el conjunto de datos también suele contener datos nominales sobre características socio-demográficas como el sexo, la ocupación, etc. No importa si los grupos (edad, sexo, experiencia laboral, tipo de ocupación, región geográfica, tamaño de la residencia, etc.) se etiquetan con números (grupo 1, grupo 2, ...) o con palabras (por ejemplo, experiencia baja, experiencia media, experiencia alta). Depende únicamente del nivel de los datos y de las operaciones matemáticas permitidas con ellos, con las que se pueden comparar los casos o las variables entre sí. Así pues, se pueden analizar los datos nominales de forma puramente cualitativa, pero también según modelos lineales generales y equivalentes. Los árboles de regresión (por ejemplo, el paquete `rpart` de R) ofrecen un método entretanto común para clasificar datos nominales, en el que el modelo y sus parámetros se desarrollan primero en un conjunto de datos de entrenamiento y esta estimación se aplica después a los datos reales de interés. Un ejemplo empírico muy utilizado es el análisis de los datos de los pasajeros del Titanic, hundido en el siglo XX, para determinar qué combinaciones de personas y características son predictores de supervivencia o muerte (por ejemplo, sexo, edad, clase socioeconómica operacionalizada por el billete, etc.). A partir de estos análisis, se pueden extraer otras conclusiones sobre los contextos sociales, los procesos a bordo durante la catástrofe del barco, etc.

13.3.2.2 Conversión de datos cualitativos en datos de frecuencia

Ejemplos típicos y conocidos de esta forma de conversión de datos son las técnicas de análisis cuantitativo de textos (Berelson, 1952; Andréu, 2011) o los análisis lexicométricos (Beaudouin, 2003; Benzécri, 2012; Tristl, Müller & Bachmann, 2015). Para interpretar textos específicos, primero se identifican las palabras clave relevantes para la pregunta de investigación y respaldadas por consideraciones teóricas. En un segundo paso, se cuantifican estas palabras clave, es decir, se contabiliza su frecuencia, para poder realizar posteriormente análisis estadísticos. De este modo, una etapa cualitativa prepara el análisis cuantitativo del texto. Por ejemplo, si un investigador quiere deducir la motivación de los aspirantes a un puesto concreto a partir de sus cartas de solicitud, puede recopilar una lista de palabras y frases críticas que se cree que están directamente relacionadas con el constructo de motivación. Esta recopilación de actitudes, opiniones y comportamientos deseables o indeseables es, sin duda, un proceso cualitativo-interpretativo. En función de

la frecuencia de dichas palabras clave o frases, se pueden categorizar y comparar aplicaciones, analizar distribuciones de características y determinar probabilidades. Y lo que es más importante, el uso del análisis exploratorio de datos puede ayudar a iluminar patrones específicos de motivación en los documentos de texto. Los estudios de seguimiento pueden incluso trabajar de forma confirmatoria sobre la base de la exploración anterior.

13.3.2.3 Conversión de los resultados del análisis cualitativo en datos CUAN

En esta aplicación de la conversión de datos para combinar datos cualitativos y cuantitativos, los resultados o resultados intermedios de un análisis cualitativo, por ejemplo, los códigos de las entrevistas o los metacódigos desarrollados a partir de ellos, se convierten en datos de frecuencia o probabilidad y, a continuación, se les utilizan para calcular los datos CUAN y se les examinan en busca de patrones específicos con ayuda de procedimientos estadísticos. Huber, Gürtler y Gento (2018) esbozan como ejemplo el análisis de entrevistas sobre características del liderazgo pedagógico (véase el capítulo 5.5.5), en el que se identificaron ocho dimensiones de comportamiento de liderazgo deseable. No fue hasta el análisis de los resultados con métodos de estadística exploratoria, en este caso análisis de conglomerados jerárquicos y escalamiento multidimensional (en 2D y 3D), que se hicieron visibles las correlaciones reveladoras entre las dimensiones individuales. Se hicieron visibles las interrelaciones entre las dimensiones individuales. Por ejemplo, se encontró una estrecha relación entre la profesionalidad de los líderes y su voluntad de implicar a los empleados de forma participativa.

13.4 CUAL y CUAN en diseños secuenciales

En el plano del diseño de estudios empíricos, Mayring (2001) ha propuesto cuatro modelos de combinación de métodos cualitativos y cuantitativos, que denomina modelo de *pre-estudio*, modelo de *generalización*, modelo de *profundización* y modelo de *triangulación*. Se pueden resumir los tres primeros modelos como macrosecuencias de combinación de métodos, mientras que la triangulación se inscribe en el ámbito de la aplicación simultánea de CUAN y CUAL. Al primer bloque de secuencias añadimos un modelo de *transformación* que falta en el de Mayring y que se aproxima a las estrategias de transformación explicadas anteriormente.

13.4.1 Macrosecuencias de la combinación de métodos

13.4.1.1 El modelo previo al estudio

Este modelo describe el enfoque según el CUAL antes solía postularse una especie de división del trabajo en la comprensión científica empírica tradicional, ... en virtud de la cual los métodos cualitativos se utilizan principalmente al principio de un proyecto de investigación en el sentido de estudios exploratorios ... Con ellos se pretende aclarar el campo de investigación, diferenciar problemas y preguntas, establecer hipótesis preliminares y comprobar la viabilidad de los métodos de investigación (Krapp, Hofer & Prell, 1982, p.130). Esta visión de *mera asistencia* a la investigación cuantitativa real *no hace justicia* a la función complementaria de los enfoques cualitativos en la investigación empírica. Sin embargo, este modelo deja claro que los estudios cuantitativos difícilmente pueden prescindir de una base cualitativa, porque es ahí donde empiezan. Y el primer paso es decisivo, ya que determina la orientación y el contenido básico. El programa de investigación Teorías subjetivas (Groeben, Wahl, Schlee & Scheele, 1988), en el que la parte cualitativa está subordinada a la cuantitativa, también puede contarse entre el modelo de estudio preliminar. Así pues, lamentablemente, una integración equitativa de los métodos y, por tanto, la resolución de la discrepancia monismo-dualismo postulada por Groeben (1986) no se lleva a cabo. Esquemáticamente, la relación en el modelo de estudio previo tiene el siguiente aspecto (Mayring, 2001, [21]):

Recordatorio 13.3: Modelo de estudio previo

Estudio preliminar cualitativo –>comprobación cuantitativa de las hipótesis –>resultados empíricos

13.4.1.2 El modelo de generalización

Según este enfoque combinado, se lleva a cabo un estudio cualitativo genuino e independiente, a cuyos resultados empíricos ya se les atribuye importancia por derecho propio. Sin embargo, de la base de datos relativamente reducida de los estudios cualitativos no se quiere o no se puede extraer – sobre todo en el contexto de la aplicación (fase de aplicación, véase la Fig. 2.1, p. 8) – ninguna conclusión trascendental para las medidas de cambio de la práctica. Esto presupone que se ha comprobado cuantitativamente la generalizabilidad de los resultados y se ha validado estadísticamente la probabilidad o la fuerza de los efectos de las nuevas medidas. Mayring (2001) pone el ejemplo de un análisis de casos, cuyos resultados se comprueban en la medida de lo posible de modo confirmatorio en un estudio de seguimiento representativo:

Recordatorio 13.4: Modelo de generalización

Estudio cualitativo –>resultados empíricos –>generalización cuantitativa

13.4.1.3 El modelo de profundización

En este modelo se invierten las etapas del diseño, es decir, a una encuesta cuantitativa y un análisis de datos diseñado para la representatividad le sigue un estudio cualitativo más pequeño, centrado en casos cuyos resultados deberían ayudar a comprender mejor la importancia de los hallazgos cuantitativos. Muchos estudios realizados en el marco del grupo de investigación sobre la juventud de Josef Held, del Departamento de Psicología de la Educación de la Universidad de Tubinga, están diseñados según este modelo (Held, 1994; Kiegelmann, Huber, Held & Ertel, 2000). En un proyecto sobre Orientaciones políticas de los trabajadores adolescentes, Held, Horn y Marvakis (1994) investigaron las razones de los adolescentes para sus orientaciones políticas. En un primer paso, los autores presentaron un cuestionario a los jóvenes; en un segundo paso, los jóvenes ayudaron a interpretar sus resultados en entrevistas de seguimiento, profundizando y concretando así los resultados cuantitativos. En este estudio, como en otro anterior de Held y Marvakis (1992), se hace visible otro efecto del modelo de profundización, su vinculación de la explicación y la aplicación, de conocimiento y la aplicación en el proceso de investigación. Una característica del procedimiento metodológico fue que el contacto con los jóvenes entrevistados no terminó con la cumplimentación de un cuestionario. A ésta le siguieron la retroalimentación y el debate de los resultados en las escuelas y empresas. Los resultados se utilizaron como estímulo para los debates en grupo y para discusiones en grupo y entrevistas en profundidad. El instrumento cuantitativo se convirtió en un medio para el cambio potencial. De este modo, el proceso de investigación se corresponde al mismo tiempo con un proceso educativo práctico, de modo que el contexto global no sólo representa una investigación académica, sino sobre todo una investigación-acción orientada a la práctica.

Recordatorio 13.5: Modelo de profundización

Estudio cuantitativo –> resultados empíricos –> profundización cualitativa

13.4.1.4 El modelo de transformación

El modelo de transformación surge de los procesos necesarios de la estadística de Bayes y describe la traducción del conocimiento cualitativo a priori en una distribución a priori bien fundamentada y adecuada al contexto (véase el capítulo 6.14.6). También se podría denominarlo *modelo de transformación del sentido común*, ya que en la estadística de Bayes, la estadística y sus operaciones lógicas se consideran una aplicación directa de los supuestos del sentido común (Jaynes, 1958). El modelo de transformación difiere del modelo de estudio previo en que el objetivo aquí no es la exploración mediante el análisis cualitativo, cuyos resultados se incorporan después *de algún modo* a los cálculos estadísticos. Se trata más bien de un arte en sí mismo para integrar la información técnica y de contenido existente y/o aún no explicada (por ejemplo, el conocimiento experto) o la información contextual de forma estructurada en una distribución a priori precisa con supuestos exactos (por ejemplo, O'Hagan e t al., 2006). Así pues, estos datos son parte integrante del análisis y no simplemente previos a un análisis numérico, pero en última instancia subordinados en términos de validez. Más bien, dependiendo del contexto, condicionan incluso los resultados en muestras más amplias, por ejemplo si las expectativas y valoraciones explícitas o los resultados empíricos anteriores son extremadamente contradictorios con los datos actuales. Están en pie de igualdad con todos los demás elementos de análisis. Aplicado a la estadística clásica, el proceso tiene el aspecto descrito por Gelman y Carlin (2014), en el sentido de que determinan el tamaño del efecto realmente efectivo para determinar un análisis de diseño (véase el capítulo 4.3.3.2) utilizando búsquedas bibliográficas y otras fuentes de información. La transformación – ya sea bayesiana o clásica – consiste en transformar el conocimiento cualitativo de forma estructurada y fundamentada en una o varias variables cuantificadoras para poder utilizarlas sistemáticamente. Debido a los supuestos muy restrictivos de la estadística clásica (véase el capítulo 4), los límites son mucho más estrechos que los de la estadística bayesiana para que tales transformaciones sean posibles en absoluto. En el caso de Bayes, el procedimiento consiste en hacer explícito el conocimiento previo y, a continuación, incorporarlo al teorema de Bayes mediante una distribución a priori o, en el caso clásico, la estimación a priori de variables relevantes como el tamaño de los efectos o el tamaño de las muestras. Esto conduce a un análisis de potencia de Neyman-Pearson o a un análisis de diseño.

Recordatorio 13.6: Modelo de transformación

Conocimiento a priori –> Teorema de Bayes (con Likelihood)

o en el caso clásico:

Conocimiento previo –> estimación de variables relevantes –> análisis de potencia o análisis de diseño.

Además, existe una gran proximidad con las estrategias de transformación introducidas anteriormente. Sin embargo, se debe distinguir el modelo de transformación de las estrategias de conversión mencionadas en que no se trata de la conversión de datos cualitativos en datos de frecuencia. Más bien, basándose en un análisis complejo de los datos cualitativos, la transformación pretende generar una distribución de probabilidad o determinar una cantidad empírica (por ejemplo, el tamaño del efecto), de modo que cada una de estas cantidades se caracterice por un contenido de información muy elevado. No se trata de convertir datos brutos en un formato analizable numéricamente, sino de una compleja actuación de integración que suele caracterizarse por un alto grado de incertidumbre y cuya cantidad resultante ya tiene una interpretación de contenido inherente. En el caso de la estadística bayesiana, se trataría del conocimiento previo sobre un fenómeno recogido a partir de diversas fuentes de información, pero completamente interconectado en sí mismo, y en la estadística clásica el equivalente sería, por ejemplo, un efecto realmente efectivo que incluyera la indicación de la altura y la dirección.

13.4.2 Aplicación simultánea de métodos cualitativos y cuantitativos

13.4.2.1 El modelo de triangulación

En este enfoque coexisten simultáneamente y en pie de igualdad diferentes enfoques metodológicos del campo de investigación. Lo importante es comparar las respuestas más o menos variables a la pregunta de investigación desde la perspectiva de los distintos métodos y sintetizar la intersección de los resultados individuales como conclusiones finales.

Recordatorio 13.7: Modelo de triangulación

Método A – *resultado global* – Método B
|
Método C

En el nivel de la triangulación de métodos, se pueden dar todas las combinaciones imaginables, es decir, en el ámbito de la investigación cualitativa, la combinación de métodos cualitativos y cuantitativos o la combinación de diferentes métodos cualitativos. No es necesario utilizarlos en una secuencia exigida por el diseño, sino simultánea y paralelamente. Un ejemplo bien fundado de esta última triangulación lo ofrecen Medina, Feliz, Domínguez y Pérez (2002, p.178), visualizado en la Figura 13.1.

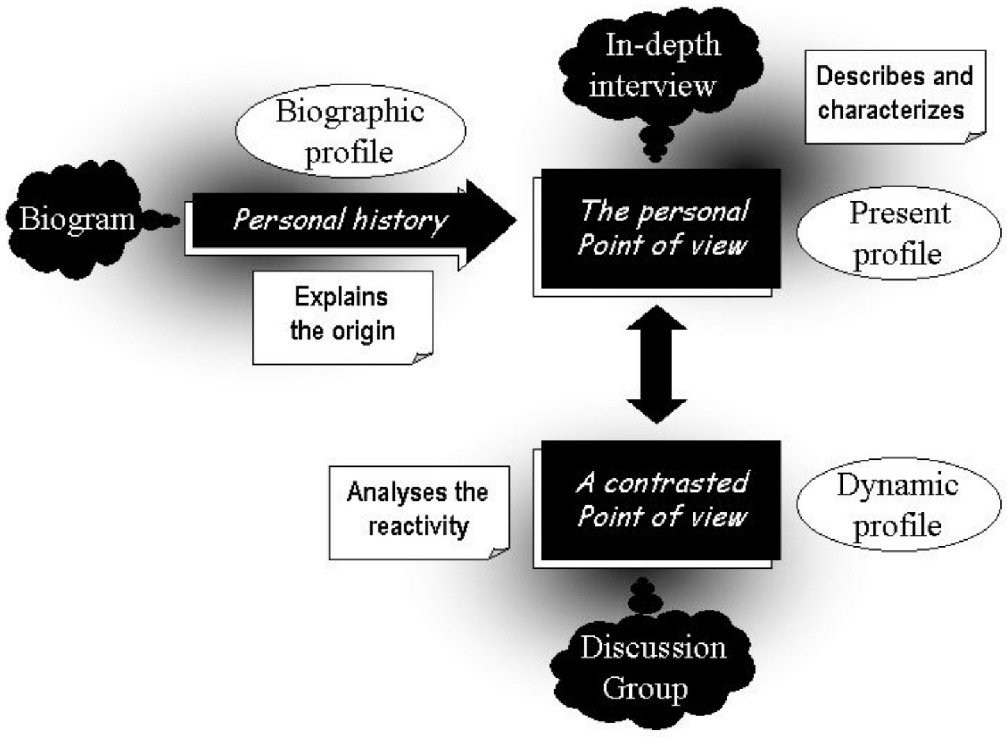


Figura 13.1: *Ejemplo de justificación de la triangulación según Medina et al. (2002)*

El *biograma* ayuda a revelar la vida personal o la historia profesional, la *entrevista en profundidad* contribuye a una comprensión diferenciada de los aspectos individuales relevantes y, por último, la *discusión en grupo* permite relacionar y comparar los puntos de vista subjetivos de cada uno de los participantes en el estudio. En la intersección de la perspectiva temporal-biográfica, la perspectiva personal-individual y la

perspectiva dinámica-social-interactiva, se encuentran respuestas equilibradas que están el abanico de las diferentes perspectivas.

Otras variantes del modelo de triangulación, por ejemplo, la combinación de las perspectivas de varios investigadores o de diferentes orientaciones teóricas, se tratan en varios artículos de Tashakkori y Teddlie (2003a), especialmente en el artículo sobre las reglas de integración de Erzberger y Kelle (2003). Sobre la relación entre los métodos mixtos y la triangulación, Gürtler y Huber (2015) ofrecen una visión general.

Por último, pasamos a un pequeño estudio de caso empírico: el papel de los pronombres personales en los debates presidenciales estadounidenses y su carácter informativo sobre el resultado de las elecciones posteriores.

13.5 Síntesis comparativa de distintos métodos de análisis

En los capítulos 9.6 (paradigma de codificación), 10.1 (análisis cuantitativo de textos) y 11.13 (análisis de secuencias), se examinó la misma carta de solicitud de plaza en un tratamiento residencial de la adicción utilizando diferentes métodos de análisis de datos. ¿Cuál fue el resultado? ¿Cómo se compararon los enfoques en este ámbito práctico? Los detalles figuran en los capítulos correspondientes. Nos adentramos en las particularidades de la aplicación de los métodos, repetimos la conclusión resumida en cada caso y terminamos con una breve comparación tabular de los métodos.

13.5.1 Paradigma de codificación

El paradigma de codificación puede utilizarse como un instrumento increíblemente flexible debido a la posibilidad de añadir codificaciones estructurales o incluso completamente diferentes a las relacionadas con el contenido, así como de combinarlas en metacódigos más abstractos. Si se añade unas hipótesis de secuencia al texto y las cuenta después, no sólo se tiene frecuencias, sino frecuencias validadas interpretativamente y comprobadas en el texto. Se trata de una categoría superior al recuento de palabras o códigos simples. Si se va más lejos, se puede pasar al análisis estadístico (exploratorio) o utilizar el análisis de implicantes. Las posibilidades son infinitas gracias a la facilidad con la que los textos codificados pueden convertirse en tablas de frecuencias. Combinaríamos el paradigma de la codificación con el análisis textual cuantitativo porque, a veces, un gráfico aquí o incluso un recuento allá o un análisis estadístico como complemento pueden dar una respuesta que no necesariamente podemos obtener así por medios exclusivamente cualitativos. Del mismo modo, puede surgirnos un interés repentino por las colocaciones, los KWIC, etc. en el contexto de las codificaciones y ahí es donde la combinación de métodos merece la pena. Hay que estar abierto a todas estas situaciones y simplemente probar.

Trabajando con el paradigma de codificación, se puede formular la conclusión preliminar del caso descrito anteriormente de tal manera que el cliente potencial se caracteriza por altas capacidades cognitivas y estratégicas con una emocionalidad simultáneamente débilmente apoyada e inherentemente inestable. Existe una alta motivación para la terapia debido a la elevada edad y a las correspondientes experiencias previas. En el plano de la acción, es de esperar que la persona tenga sus propios puntos de vista sobre la aplicación e interpretación de los procedimientos y contenidos en la terapia cotidiana, el concepto y las las normas institucionales necesarias.

13.5.2 Análisis cuantitativo de textos

El análisis cuantitativo de textos resultó ser el procedimiento más difícil en este caso, ya que simplemente había muy pocos datos y muchas menciones individuales, lo que ya era visible en el paradigma de codificación. Contar las menciones individuales es aburrido, pero también podemos leer fácilmente el texto ciertamente muy corto. Un análisis cuantitativo se nutre de la comparación de similitudes y diferencias en términos de frecuencias u otros valores numéricos de referencia. Aunque existen muchas posibilidades para

analizar textos en R, aquí sólo se han podido probar en casos singulares, que son demasiado pocos para sacar conclusiones sustanciales. La preparación inicial del material de datos es importante, si se elimina del conjunto de datos qué y cómo. Pero este procedimiento sólo tiene sentido para las comparaciones, de lo contrario no es necesario.

No obstante, se pueden formular hipótesis iniciales sobre la persona, pero aún no una estructura de caso: la perspectiva en primera persona se encuentra con la otra, la contraparte. Relevantes -según la frecuencia como criterio – son los cinco términos yo, mí, vosotros, eso y éstos. Todo lo demás son menciones individuales, lo que pone de relieve una cierta variabilidad en la expresión a pesar de la brevedad de la carta.

Las aptitudes cognitivas están presentes en el candidato y podrían servir de recurso en el futuro. La emocionalidad o la motivación no están en el punto de mira y suelen entrar en juego de forma encubierta. Si se analiza la relación entre Yo y Vosotros, se observa el estilo de tutear y la ignorancia asociada de la relación formal. Hemos interpretado esto como posibles problemas con reglas durante la terapia. Se nota que una simple motivación de la terapia no es expresado, sino que, en realidad, el texto está atascado en el nivel del yo y el tú. En la comparación con el paradigma de codificación y el análisis de secuencia, el análisis es mucho más modesto – por lo que necesitamos más espacio para presentarlo aquí. Lo que falta es la integración de los elementos individuales.

Sin embargo, con un texto más largo y más aún con varios textos, esperamos que esta desventaja pueda compensarse. En ese caso, se requiere una buena estrategia de análisis para saber dónde hay que centrarse. Por ejemplo, de forma similar al paradigma de codificación, se podría crear un diccionario para asignar ciertas palabras dentro y fuera de su contexto a algún tipo de codificación. Entonces, esta clase de procedimiento se vuelve muy potente y esto crece con el aumento del material de datos. De forma comparable a las hipótesis de secuencia, se pueden establecer relaciones entre y dentro de textos o partes de textos, lo que sin duda da lugar a perspectivas interesantes.

Lo que se omitió deliberadamente en el presente análisis es la combinación del análisis cuantitativo de textos con las técnicas de interpretación de los otros dos métodos. Sin embargo, lo recomendamos encarecidamente a efectos prácticos. No se ha hecho con fines demostrativos. Ahora se podrían analizar los cinco términos de modo interpretativo y examinar el contexto – sólo por mencionar una posible estrategia de combinación de métodos.

13.5.3 Análisis de secuencia

En la práctica, el análisis de secuencia resulta comparable al trabajo con la estadística bayesiana. Trabaja con lo que encuentra y aún así puede producir conclusiones coherentes. La cantidad de material de datos es en cierto modo insignificante porque no se trata de eso. Por lo tanto, aparte de los aspectos formales como el asunto y el saludo, media frase era suficiente para el análisis. En general, sólo se utilizó una frase para la falsificación. En la práctica, por supuesto, no es así como se trabaja, pero demuestra que ciertamente se puede llegar lejos con poco material y un nivel de esfuerzo correspondientemente alto. El análisis de secuencia es autosuficiente y, en principio, no necesita ampliarse con el paradigma de la codificación ni con el análisis cuantitativo de textos, aunque, en nuestra opinión, estos dos métodos pueden complementarse muy bien. Pero nada excluye posibles combinaciones con el análisis secuencial. Nos limitamos a subrayar la autonomía del procedimiento.

La hipótesis preliminar de la estructura del caso, que no puede falsarse en el marco del análisis, es: emocionalidad débilmente apoyada con capacidades cognitivas pronunciadas y una fuerte tendencia a evitar el encuentro del yo como agente.

13.5.4 Comparación de los enfoques

En la tabla 13.2 se resumen los enfoques según el enfoque que pudo cubrir cada uno de los temas mencionados. Las notas indican, por ejemplo, los casos donde no se examinaron suficientes datos o no se disponía de ellos, o no se alcanzó un resultado o un resultado no podía deducirse directamente pero estaba implícitamente presente. El análisis resumido del paradigma de codificación y el análisis secuencial se solapan más o menos completamente. El análisis cuantitativo de textos es naturalmente diferente, ya que

trabaja a través de la muy concreta orientación a las palabras y menos a las interpretaciones de las cosas y, en consecuencia, las conclusiones preliminares están en un lenguaje distinto.

- Los tres enfoques coinciden en que el cliente potencial tiene altas capacidades cognitivas. El análisis cuantitativo del texto también subraya que en el texto se pueden encontrar estructuras de argumentación y un enfoque concreto. Esto también se podría deducir del paradigma de codificación y del análisis de secuencia, pero no era el argumento central.
- El paradigma de codificación y el análisis de secuencia coinciden en que las cogniciones fuertes van acompañadas de una emocionalidad débilmente apoyada. El análisis cuantitativo de textos tuvo problemas para elaborar la hipótesis de la emocionalidad débilmente apoyada, sino más bien con la escasa cantidad de datos.
- El análisis cuantitativo de texto elabora como perspectiva central yo-contra-tú, lo que en los otros dos procedimientos se toma más bien como evitación del encuentro concreto y, por tanto, claramente más acertado.
- La elevada edad y la motivación terapéutica que la acompaña no necesitan ningún análisis especial, ya que esto casi se puede deducir del sentido común y es simplemente obvio en el texto.
- El potencial del cliente para cuestionar las normas o no adherirse a ellas o interpretarlas según su propio punto de vista sale a la luz precisamente a través del paradigma de codificación. Sin embargo, el análisis de secuencia sólo analizaba media frase. Con el análisis de toda la carta, cabría esperaríamos que también se encontrara esa figura, bien justificada. El análisis textual cuantitativo puede puntuar de nuevo con la figura de la perspectiva yo-tú.

Tabla 13.2: *Comparación de métodos de análisis de datos (carta de solicitud para terapia de adicciones)*

	Cogniciones fuertes	Emocionalidad débilmente apoyada	Evitación de encuentros	Yo vs. Tu Perspectiva	Motivación-alta	probalem. infracciones de normas	Estrategia/ Argumentaciones
Paradigma de Codificación	×	×	×	–	×	×	×[4]
Análisis cuantitat. de textos	×[1]	_[2]	–	×	×	×	×
Análisis de secuencia	×	×	×	–	×	?[3]	×[4]

[1]Deducción indirecta (conceptos múltiples)
[2]Necesita más datos
[3] Deductable con análisis de más datos
[4]implicita en la figura de "cogniciones"

En definitiva, no parece que un método sea realmente superior al otro a largo plazo. Todos tienen el potencial de llegar a conclusiones comparables. Es probable que el esfuerzo sea el mismo en cada caso y que siempre haya que estimarlo en un nivel alto. No hay que engañarse pensando que el análisis cuantitativo de textos es superficialmente inferior. Por su modo de funcionamiento requiere condiciones marco diferentes y no tiene como objetivo el analsisis de unas pocas palabras, sino que crece con sus tareas, es decir, con la cantidad de datos de que se dispone. Partimos de la base de un tratamiento inteligente de los datos y no de un enfoque ciego y sin teoría de la escopeta, como suele sugerir el término minería de datos.

Las diferencias prácticas entre los métodos surgen del hecho de que el análisis cuantitativo de textos requiere una cierta cantidad de datos, el análisis de secuencias se arregla con unas pocas palabras en el peor de los casos, y el paradigma de codificación permite estrategias de búsqueda diversas y anidadas, que, sin embargo, también requieren ciertas cantidades de datos para las comparaciones. Una vez más, la idea de iniciar algo así como una disputa de paradigmas sobre la superioridad de uno de estos métodos es, en nuestra

opinión, una burda travesura. En la combinación de los métodos, sin embargo, no queda realmente nada que desear.

Como conclusión de los Métodos Mixtos, pasamos ahora a la cuestión de cómo llegar a convertirse en presidente de los EE.UU. Se presenta una integración esquemática de los métodos.

13.6 Esquema de la integración de un método – ¿Cómo llegar a ser presidente en EE.UU.?

13.6.1 Pregunta

Como último ejemplo de integración de métodos, nos basamos en el estudio de caso detallado del capítulo 6.14.3 o 6.15.3 sobre una teoría de la ciencia política. El supuesto examinado es que en los debates presidenciales previos a unas elecciones – aquí: en EE.UU. – el aspirante tiene que conseguir que *la nación le respalde* y, por tanto, tiende a ser mayoritario. y, por tanto, tiende a hablar en la forma *nosotros* y menos en la forma *egocéntrica yo*. El objetivo es conseguir el apoyo de la nación y reemplazar así al titular. Si, por el contrario, un titular entra en un duelo discursivo, debe *referirse a sí mismo y a los éxitos logrados en la pasada legislatura*, demostrando así que sólo con él está asegurado el bienestar de la nación. Esto habla en favor de una estrategia orientada hacia la primera persona. Esto se puede comprobar utilizando como base de datos las transcripciones existentes de los debates presidenciales estadounidenses. Éstas pueden encontrarse continuamente a partir de 1988 en la Commission on Presidential Debates (1987). En los análisis de los capítulos 6.14.3 y 6.15.3, la atención se centró en el duelo entre los En los análisis de los capítulos 6.14.3 y 6.15.3, la atención se centró en el duelo de 2004 entre *el titular y el aspirante*: George W. Bush vs. John F. Kerry.

Se realizaron diversos análisis estadísticos utilizando diferentes procedimientos: clásico, bayesiano con simulación MCMC y como prueba t bayesiana exacta basada en una solución analítica del problema Behrens-Fisher de Bretthorst (1993). Damos por sentada la lectura de esta sección y ahora nos centramos más en la parte cualitativa, es decir, en cómo un análisis cualitativo puede aumentar sustancialmente el contenido informativo de los análisis estadísticos. Esto requiere tiempo y mucho esfuerzo.

Todo estudio empírico necesita una pregunta de investigación sustantiva. Una pregunta de investigación plausible en nuestro caso es:

Caso 13.1: La lengua y la campaña electoral estadounidense

¿Qué énfasis en términos de elección lingüística y autodramatización pone un candidato u orador en campaña en su discurso para que después se le considere un ganador del discurso y, en consecuencia, aumente las posibilidades de victoria de su propia campaña?

Se puede analizar la tesis a partir de unas pocas palabras clave para cada duelo de discursos individualmente. Metodológicamente, se lleva a cabo un análisis de contenido cuantitativo en AQUAD 7 (Huber & Gürtler, 2012) y posteriormente se comprueba los resultados en R con métodos estadísticos inferenciales. De ello se ocupan los análisis de los capítulos 6.14.3 y 6.15.3. Una codificación cualitativa de los pasajes de texto relevantes en el material original con respecto a los polos yo frente a nación, que incluya específicamente el contexto de las palabras clave, sería más elaborada y significativamente más exacta. Esto puede hacerse mediante hipótesis y codificación de secuencia (véase el cap. 9.5), como por ejemplo que un orador intente anteponer sus propios éxitos o que la nación le respalde, y no solo para cuando un orador tiene algo que mostrar, sino igualmente cuando puede que no haya mucho. Desde Donald Trump y la campaña electoral de 2016, ha habido una estrategia consistente en no argumentar realmente los éxitos ni

integrarse en términos de apoyo de la nación, sino separar populistamente y crear miedo en la nación con la esperanza de que la división y la inseguridad resultantes ayudarán al propio éxito electoral. Que estas sean entonces las razones reales es otra cosa - ¿entonces ganó Donald Trump o perdió Hillary Clinton en 2016? Del mismo modo, dejamos de lado la enorme influencia de las redes sociales en la campaña electoral. Sería interesante comparar el debate de 2016 con otros debates presidenciales que siguieron un patrón más tradicional de intentar convencer con argumentos. No significa, sin embargo, que esto se haya aplicado siempre de forma coherente y que no se hayan utilizado también elementos del estilo altamente agresivo de Donald Trump. Desde Donald Trump, sin embargo, los debates han perdido el supuesto prioritario de que se basan realmente en argumentos. Ha quedado claro que, en tiempos de autorrepresentación e influencia mediática, los argumentos son, en el mejor de los casos, una cuestión secundaria.

Nos hemos abstenido deliberadamente de analizar los debates de Trump y Clinton, ya que éstos abren un nuevo tipo y difieren cualitativamente de los debates anteriores.

13.6.2 Preparación y codificación de los datos

Los datos se leen en AQUAD 7 (Huber & Gürtler, 2012, véase el capítulo A) y primero se codifican o se definen y cuentan las palabras clave. Los titulares se codifican de forma diferente con códigos de orador que los aspirantes. Esto tiene en cuenta el caso en que dos aspirantes se enfrentan y ninguno de ellos es un titular, de modo que el titular no puede ser acusado de ningún error o, por el contrario, ningún titular está presente para impugnar falta de experiencia al aspirante. Por lo tanto, se podría suponer que dos aspirantes se reúnen de manera diferente que un aspirante y un titular. En el caso de dos aspirantes, sería razonable suponer que *los éxitos personales* deben conceptualizarse de forma diferente, ya que hasta ahora no es el cargo del presidente el que está en el trasfondo como legitimación, independientemente de si se han logrado éxitos objetivamente mensurables (por ejemplo, evitar la recesión económica, luchar contra las enfermedades y la pobreza, etc.). Ambos aspirantes necesitan que la nación les respalde detrás de ellos – y así se puede examinar si se utilizan estrategias de comunicación justas (por ejemplo, cuestiones objetivas) o injustas (por ejemplo, devaluaciones personales de la otra persona, mentiras, etc.) o en qué medida se utiliza cada una de ellas. Para una prueba estadística, resulta una constelación diferente cuando se enfrentan dos aspirantes que cuando uno de ellos es un titular. Así, cabría esperar que, según el tipo de duelo discursivo, distintos factores son relevantes para representarse a uno mismo como presidente potencial que cuando se trata de candidato frente a presidente en funciones. Si en este punto se recurre a información cualitativa contextual adicional, por ejemplo sobre la forma en que Donald Trump se comunicó durante la campaña electoral dentro del Partido Republicano o posteriormente contra Hillary Clinton, ya sería necesario realizar ajustes a la teoría existente, como se ha mencionado anteriormente. No se trata entonces de yo o nosotros/nación, sino del peyorativo tú. En los capítulos 6.14.3 y 6.15.3, omitimos deliberadamente esta información contextual relevante y nos concentramos en las simples palabras clave del caso. Esto dio la impresión, apoyada por las estadísticas, de que Bush y Kerry practicaban estrategias muy similares.

La amplia inclusión de información contextual requeriría un análisis cualitativo muy detallado para identificar con claridad los mecanismos comunicativos que operan en el texto en cada caso, lo que excede con mucho el alcance del libro. Además de la codificación pegajosa de los pasajes del texto, la formulación de hipótesis sobre el modo en que y las formas de utilizar los componentes estratégicos para presentarse como mejor o para o para devaluar al adversario y calificarlo de incompetente o indigno de confianza. Éstas pueden ponerse a prueba mediante hipótesis secuenciales con el fin de registrar el resultado como un código independiente.

13.6.3 Estrategia de análisis de datos y relación con la pregunta de investigación

Independientemente de cómo se codifiquen e independientemente de lo complejos que sean estos códigos ampliados mediante hipótesis cualitativas, el resultado son recuentos de categorías y códigos más o menos complejos. El resultado es, por tanto, un conjunto muy agregado de recuentos, cuyo contenido no corresponde a secciones de texto simplemente codificadas, que a menudo se parecen más bien a simples resúmenes, sino a hipótesis ya probadas en el texto. Todos los resultados estadísticos se refieren ahora a esta información condensada, que va mucho más allá de un simple análisis de contenido cuantitativo y corresponde a la reivindicación de los métodos mixtos. Un simple análisis estadístico no podría alcanzar

fácilmente este nivel de densidad de información subyacente si sólo se dispusiera de códigos sencillos o incluso si sólo se contaran las palabras clave, como aquí en el caso de Bush frente a Kerry. Necesita ambas cosas: procesar la información y relacionar esta información según unos criterios definidos.

Las estadísticas pueden ser tan elaboradas como se desee. Por ejemplo, modelos lineales (jerárquicos) para tener en cuenta las condiciones específicas de cada duelo de discursos, la cuestión de si un orador es titular o candidato, etc., en el marco de las llamadas sectas aleatorias. De este modo, los efectos temporales y secuenciales pueden modelizarse adecuadamente dentro de los tres duelos de oradores y a lo largo del tiempo entre campañas electorales. Los predictores externos adicionales pueden ampliar el modelo estadístico – ya sea la afiliación a un partido político, el entorno socioeconómico, la edad, el tamaño corporal, el coeficiente intelectual, etc. – y pueden utilizarse para modelizar los efectos de los tres duelos electorales. Por supuesto, esto requiere una justificación sustantiva de por qué un predictor debe incluirse en el modelo.

Claro, sería beneficioso realizar un estudio completo de todas las transcripciones disponibles de los debates presidenciales. Ficticiamente, esto tendría el siguiente aspecto: La atención no se centra sólo en las palabras clave, sino en códigos muy complejos de captar, de modo que no sólo palabras sueltas, sino también las secuencias de palabras, los enunciados o los pasajes del discurso categorizados y contextualizados de forma significativa forman el objeto contable de análisis. De este modo, la cuestión general planteada anteriormente se puede describir de forma mucho más realista, ya que no se trata sólo de palabras clave, sino de un uso de términos y frases sensibles al contexto. Las palabras clave y las expresiones idiomáticas tendrían que traducirse de acuerdo con el espíritu de la época para garantizar que los análisis no surjan del significado actual de la lengua, sino que se ajusten al contexto en el que se originaron. Esto requeriría una tabla de traducción exhaustiva que incluyera qué modismos y términos corresponden a qué otros de diferentes épocas, para que hacer el proceso lo más transparente posible. Un análisis de este tipo es multiprofesional, ya que podrían participar tanto politólogos y analistas lingüísticos como historiadores, metodólogos cualitativos o estadísticos.

Como ampliación, es concebible evaluar o analizar cualitativamente todos los duelos de discursos individualmente y con independencia de los participantes mediante categorías adicionales que deben ser explicadas y seguramente también por expertos externos, y utilizar después el análisis de implicantes (véase el capítulo 12) para la tipificación de las transcripciones de los discursos. Esto no requiere necesariamente sólo categorías muy agregadas, ya que el procedimiento no excluye explícitamente ni prefiere una mezcla de diferentes niveles de complejidad. Esto depende de la pregunta de investigación y de si tiene sentido. Lo mismo se aplica al análisis estadístico. Sin embargo, son preferibles las categorías complejas con un alto contenido de información, ya que conservan de forma natural los análisis que desembocaron en ellas en análisis posteriores y el significado de los elementos de análisis es sencillamente mayor.

Volviendo al análisis de implicantes (la minimización booleana), éste puede responder a la pregunta de qué factores lógicamente causan el éxito o el fracaso electoral. El criterio es entonces el éxito electoral en términos positivos y negativos, con el fin de encontrar configuraciones para el éxito electoral en una elección presidencial estadounidense. La minimización booleana permite responder a esta pregunta siguiendo una lógica binaria. A su vez, los modelos lineales permiten estimar la contribución relativa de los predictores o de las interacciones de los predictores. Creatividad no tiene límites.

13.6.4 Integración de datos

No hemos realizado los análisis bastante complejos descritos anteriormente. Sin embargo, serían fácilmente posibles en el marco de un trabajo exhaustivo. Pero el procedimiento esbozado describe lo que se puede realizar cuando se aplica una combinación de metodología cualitativa y cuantitativa. No sólo se puede condensar increíblemente la información, sino que se puede combinarla fácilmente con otros factores en el marco de cualquier – y queremos de hecho decir cualquier – procedimiento de análisis de datos. Esto trasciende la posición del análisis cualitativo como trabajo preliminar o proceso generador de hipótesis, como ciertamente se sigue describiendolo todavía en algunos libros de texto. La secuencia temporal resulta de la secuencia natural de que primero hay que condensar la información antes de relacionarla entre sí y con los demás. La utilización simultánea y equivalente de la estadística y la minimización booleana en las variantes binaria y de lógica difusa, respectivamente, ofrece precisamente el contraste necesario para poner de manifiesto los extremos. Por un lado, la minimización booleana polariza y reduce a lo esencial, por otro lado la estadística recoge la imprecisión que ofrece el mundo y que se pierde con una lógica binaria.

Tarea 13.1: Estudio de caso de métodos mixtos e integración de datos

En los capítulos 9.6, 10.1 y 11.13 hemos utilizado el mismo material, una carta de solicitud de una retirada psiquiátrica para una plaza en un tratamiento hospitalario de adicciones, utilizando diferentes métodos de análisis de datos: paradigma de codificación, análisis cuantitativo de texto y secuencia Análisis secuencial.

La tarea del lector consiste ahora en integrar los resultados allí encontrados y decidir si los resultados apuntan a lo mismo en términos de contenido o si permiten interpretaciones e interpretaciones completamente diferentes.

Ambas cosas requieren un análisis cualitativo posterior, para integrar los resultados del análisis (= compresión de la información), minimización booleana y estadística. Esto casi se parece a un metaanálisis y el análisis cualitativo se presta a ello porque de lo contrario se necesitarían traducciones adicionales y muy complejas para transformar los resultados cualitativos en valores numéricos de salida para un metaanálisis estadístico o para mejorar la imprecisión de los resultados estadísticos con el fin de examinarlos metaanalíticamente mediante una nueva minimización booleana. El análisis cualitativo deja los datos como están y se limita a registrar sus enunciados para posteriores comparaciones.

Partimos de la base de que *todo análisis* – cualitativo, cuantitativo, binario o de lógica difusa – se basa fundamentalmente en la tecnología cualitativa en cuanto a su transformación en una respuesta utilizable a una pregunta inicial, es decir, está sujeto a un proceso de traducción en varias etapas (véase también Gigerenzer, 1981; Gürtler, 2005). Por lo tanto, para nosotros está fuera de lugar que la integración de diferentes resultados analíticos tenga que ir exactamente por este camino - a saber, la traducción al respectivo al otro lenguaje metodológico respectivo con el fin de llevar dialécticamente los resultados disponibles a un nivel superior e integrarlos allí. para responder a la pregunta capciosa.

13.6.5 Interpretación metodológica

Las siguientes interpretaciones se centran en la interpretación metodológica y el valor añadido mediante el uso combinado de métodos. No se lleva a cabo una interpretación del contenido a nivel de teorías políticas o de acción comunicativa; eso debería ser tarea del lector políticamente interesado.

En primer lugar, cabe señalar que sólo se puede contar cuando está claro qué es lo que hay que contar. Lo mismo ocurre cuando se trata de estimar probabilidades y no de contar. Esto requiere un cierto análisis cualitativo para fundamentar el punto de partida de cualquier estadística. Que las palabras clave individuales se definan como una unidad o que se utilicen pasajes complejos de texto para formar categorías mucho más sustanciales es irrelevante y no implica ninguna diferencia estructural. Independientemente de la complejidad del análisis cualitativo, su influencia en el nivel numérico es tan grande que no se puede hablar en ningún caso de un estudio preliminar. Más bien, la estadística es una consecuencia lógica del análisis cualitativo. A la inversa, ni siquiera los análisis cualitativos bien fundados no pueden responder a la cuestión de las dependencias numéricas.

En este sentido, ambos arsenales de métodos son aquí inseparables y se complementan o fusionan a la perfección, aunque es posible una gran flexibilidad dentro de los enfoques. En lo que respecta a la integración, ningún enfoque es superior o está subordinado al otro. Se trata más bien de utilizar toda la información disponible. Así, el análisis cualitativo de datos puede ser tanto simple como complejo, del mismo modo que la estadística puede ser clásica o bayesiana. Además, son posibles diferentes métodos analíticos, cada uno de los cuales es capaz de abordar diferentes cuestiones y grados de complejidad.

En lo que sigue es importante que no tratemos simplemente con resultados binarios como hipótesis *aceptada frente a rechazada*, sino con rangos de resultados provistos cada uno de ellos de diferentes grados de incertidumbre. No sólo se debe entender la incertidumbre desde el punto de vista estadístico, sino que también afecta a los resultados de los análisis cualitativos de datos, que tampoco son cien por cien

inequívocos, pero cuyas interpretaciones, al igual que las estadísticas, pueden y deben tener un alto grado de verosimilitud. Esto hace que merezca la pena hablar de un e*spacio de resultados* y no de un resultado singular, ya que, además de las explicaciones de los capítulos 1.2 y 1.3 sobre la filosofía de la ciencia, se aplica un criterio de verdad relevante en el que coexisten muchas explicaciones diferentes paralelas entre sí. Por supuesto, sería posible, con la incertidumbre adecuada, concentrarse en una única solución favorecida que pudiera prevalecer sobre las interpretaciones alternativas. Sin embargo, nuevos datos podrían y posiblemente cambiarían esta solución. Más prometedora es la orientación inmediata hacia la complejidad, para integrar sintéticamente en lugar de eliminar los modelos que compiten entre sí. Al final, probablemente sea más fácil asumir desde el principio el obstáculo de trabajar con varios enfoques explicativos en paralelo y ver si y cómo cambian o se comportan en qué condiciones contextuales, para luego integrarlos en un modelo complejo. En el proceso, se examina qué tesis permanecen constantes, cuáles desaparecen, cuáles cambian y cuáles se añaden, en función del contexto.

Este procedimiento ya se conoce del análisis de secuencia, es decir, de parte de texto a parte de texto en la generación de la hipótesis de estructura de caso propuesta. El procedimiento esbozado se aplica por igual al trabajo cuantitativo y cualitativo y a su combinación en cualquier caso. El ejemplo empírico de los debates presidenciales pretende ilustrar el potencial de un enfoque tan complejo. En el estudio del caso, la atención se centró en la elección de los términos estudiados, es decir, las palabras clave: yo, nosotros y nación. Al principio de un análisis más complejo, pues, está la observación de que estos términos no son unívocos y, por tanto, cualquier conclusión que asuma la univocidad de los términos, y por tanto el análisis contable de palabras clave singulares, debe estar muy viciada. Esto viene a demostrar que el análisis de contenido ciego, cuantitativo y descontextualizado funciona de forma inadecuada. Así, en el caso de Clinton frente a Trump en 2016, se puede demostrar que Trump utiliza el término nosotros en referencia a sí mismo y no como un nosotros real en el sentido de formar parte de algo más grande, lo que a su vez llevaría a la nación. Es muy posible, sin embargo, que otros candidatos antes que él se refirieran realmente a nosotros con el término nosotros y se sintieran parte de la nación y no al revés en el pluralis majestatis o incluso más según el absolutista L'état c'est moi. Trump actúa así en el sentido de extender su propio ego al Estado, de modo que para él el concepto de nosotros debe reinterpretarse de forma casuística. Sin duda, esto podría demostrarse fácilmente en el curso de un análisis de secuencia fina a nivel textual. No se trata de los demás, sino exclusivamente de uno mismo en la generalización (es decir, en el sentido de clonación) de sí mismo a muchos otros. En consecuencia, los demás incluidos instrumentalmente sólo si encajan en el propio esquema y, por lo demás, eliminados del propio espacio de acontecimientos, ya que – según la tesis – nada existe fuera de uno mismo. Con esta hipótesis se pueden explicar los diferentes usos del término nosotros en los demás debates de elecciones anteriores, si y en qué interpretación analítica de los términos investigados y codificarlos en consecuencia. No se trata, pues, de interpretaciones individuales, sino de interpretaciones sensibles al contexto, pues incluso podrían diferir en función del debate, aunque esto no parezca muy plausible, ya que las personas actúan de forma bastante constante y cabe suponer que la retórica tiene una cierta estabilidad. Sin duda hay otros niveles de significado del nosotros que habría que explicar con más detalle. A partir de ahí, se forman otras categorías adicionales que proporcionan una base aún más sustancial para un posterior análisis cuantitativo del contenido. Este tipo de combinación de métodos parece prometedora y mejor que realizarla *a ciegas* y sin interpretación contextual. Y un procedimiento de este tipo puede ayudar, sobre todo en casos complejos, a mantener en paralelo diferentes modelos prometedores y explorar su ámbito de aplicación, en lugar de ponerlos a prueba entre sí y considerar un *único enfoque verdadero* como explicación válida del comportamiento de los candidatos presidenciales.

De hecho, las hipótesis pueden examinarse en relación con el texto, las referencias hechas, etc., antes de contarlas. La base de cada estadística siguiente se ve así fuertemente enriquecida en términos de contenido, condensada y sustancial, puesto que ya se han realizado pruebas cualitativas. Así pues, los enunciados estadísticos son también de naturaleza más sustancial y, por lo tanto, pueden diferir del caso no complejo.

Tarea 13.2: Campaña electoral estadounidense - análisis complementario

La tarea para el lector interesado consistiría en realizar uno de los análisis mencionados, por ejemplo, sobre un debate anterior. Alternativamente, se podría examinar cómo las tesis individuales, aún inconexas, sobre el uso de palabras o frases clave pueden integrarse en un sentido dialéctico y qué predicciones contrapuestas se pueden hacer para futuras elecciones en los próximos años, o retrospectivamente en relación con elecciones anteriores y sus resultados.

Este contexto se puede contrastar fácilmente con la realidad, es decir, quién ganó las elecciones en una fase temprana y qué modelo es probable que haya tenido éxito y qué cambios cabe esperar en el futuro. Si hay debates en otros países, también se podrían analizarlos. Las posibilidades de análisis son ilimitadas.

En resumen, el estudio de caso empírico se sitúa en el ámbito del análisis combinado de datos. A partir del análisis de contenido cuantitativo (enumeración pura), se puede derivar una hipótesis estadística sobre la distribución de las palabras clave (categorías). Esto se complementa con observaciones a nivel cualitativo – por ejemplo, que Trump no habla de la nación, sino que sólo habla de sí mismo y se refiere exactamente a eso. Para ello, se pueden consultar fuentes externas, como otros discursos y apariciones de Trump, con el fin de examinar esta tesis más de cerca y verificarla cualitativamente. Este tipo de argumentación cualitativamente diferente de Trump, en contraste con candidatos anteriores, por ejemplo, puede llevar entonces a un reanálisis retrospectivo y a una reinterpretación del contenido de debates anteriores a la luz de estos nuevos puntos de vista.

Todo esto habla claramente *en contra* de un recuento ciego de palabras clave y de un análisis de contenido cuantitativo irreflexivo que pega literalmente a la palabra y no al significado y se limita al recuento descontextualizado. A partir de los resultados descritos se puede derivar una tipología y combinarla con el éxito electoral. Teóricamente, se podrían desarrollar partes de una estrategia de campaña electoral o se podría evaluar críticamente la validez del modelo en otros países democráticos.

Bibliografía

[Laotse], Laozî (2007). Dàodéjing [Tao Te King]. München: Heinrich Hugendubel Verlag [Online: Projekt Gutenberg http : / / projekt . gutenberg . de]. url: http : / / gutenberg . spiegel.de/buch/tao-te-king-195/4 (visited on 15. 06. 2019).

3851283 https://stats.stackexchange.com/users/63568/user3851283, user (12. Dez. 2014). What's a real-world example of överfitting? Cross Validated. url: https://stats. stackexchange.com/q/128616 (visited on 29. 05. 2019).

A complete guide to the Bayes Factor test (2016-09-13). Techn. Ber. Defazio, Aaron. url: https / / www . aarondefazio . com / adefazio - bayesfactor - guide . pdf (visited on 24. 05. 2019).

Actualitix (10. Jan. 2016). Tee — Exportländer ($). Karte | Rangordnung der Statistiken | Datentabelle. url: https://de.actualitix.com/land/wld/tee- exportlander.php (visited on 02. 05. 2019).

Aebli, Hans (1980). Denken. Das Ordnen des Tuns. Band I: Kognitive Aspekte der Handlungstheorie. Stuttgart: Klett-Cotta.

Agresti, Alan & Barbara Finlay, (Eds.). (1997). *Statistical Methods for Social Sciences*. Prentice Hall.

Aitkin, Murray (1997). *The calibration of P-values, posterior Bayes factors and the AIC from the posterior distribution of the likelihood*. Statistics and Computing. https : / / www . newscientist . com / article / dn20447 - journal - rejects - studies - contradicting-precognition.

Akaike, Hirotugu (1983). Information theory and an extension of the maximum likelihood principle. 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2-8,1971, Budapest. Hrsg. von B.N. Petrov & F. Csáki. Republished in Kotz, S. & Johnson, N.L. (Editors). (1992). *Breakthroughs in Statistics, I*, Springer-Verlag, pp. 610– 624. Budapest: Akadémiai Kiadó, pp. 267-281.

Albert, Jim (2007). *Bayesian Computation with R*. Dordrecht: Springer.

Alder, B.J. & T.E. Wainwright (1959). "Studies in Molecular Dynamics. I. General Method". Journal of Chemical Physics 31(2). https : / / pdfs . semanticscholar . org / 9fdf / b307864a0893fb0deb7be0b6c701787a092a . pdf, S. 459–466. (Visited on 05. 06. 2019).

Aldhous, Peter (2011-05-05). Journal rejects studies contradicting precognition. *New Scientist* [The Daily Newsletter]. https://www.newscientist.com/article/dn20447- journal-rejects-studies-contradicting-precognition.

Aldrich, John (1997). R.A. Fisher and the Making of Maximum Likelihood 1912–1922. *Statistical Science 12*(3). https://projecteuclid.org/download/pdf_1/euclid.ss/ 1030037906, S. 162–176.

Aldrich, John (2000). Fisher's inverse probability of 1930 *International Statistical Review 68*(2), 155–172. https://www.jstor.org/stable/1403666.

Alea Riera, V. et al. (2014). Guía para el análisis estadístico con R Commander. *Publicacions i Edicions de la Universitat de Barcelona*. Barcelona: Universitat de Barcelona.

Allison, Paul (5. März 2015). Imputation by Predictive Mean Matching: Promise and Peril. *Statistical Horizons*. url: https://statisticalhorizons.com/predictive-mean-matching (visited on 05. 03. 2019).

AMR (16. Feb. 2018). Simple Fast Exploratory Data Analysis in R with DataExplorer Package. *Towards Data Science* [Blog]. url: https : / / towardsdatascience . com / simple - fast - exploratory - data - analysis - in - r - with - dataexplorer - package - e055348d9619.

Anderson, Edgar (1935). The irises of the Gaspé Peninsula. *Bulletin of the American Iris Society, 59*, 2–5.

Anderson, Edgar (1936). The species problem in Iris. *Annals of the Missouri Botanical Garden, 23*(3), 457–509. http://biostor.org/reference/11559, S..

Anderson, Valerie (2017). Criteria for Evaluating Qualitative Research. *Human Resource Development Quarterly 28*(2), 125–133.

Anderssen-Reuster, Ulrike, Hrsg. (2007). *Achtsamkeit in Psychotherapie & Psychosomatik. Haltung und Methode*. Stuttgart: Schattauer.

Andréu Abela, J. (2011). Las técnicas de análisis de contenido: Una revisión actualizada. Granada: Universitat de Granada.

Andrews, D.W. (1991). Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation. *Econometrica 59*, 817–858.

APA (1996-03). *P-values under question. American Statistical Association issues statement on proper role of significance testing in research.* Techn. Rep. https : / / www . apa . org / science / about/psa/2016/03/p-values.

APPS, Clinical Research (2017). *RRApp Robust Randomization App* [version 3.0.I]. developed by Chengcheng Tu, MPH, under the mentorship of Dr. Emma Benn. url: http : / / clinicalresearch-apps.com/RRApp.html (visited on 27. 06. 2019).

APS — Association for Psychological Science (o. D.). Registered Replication Reports. Instructions for Authors: https://www.psychologicalscience.org/publications/ampps/ registered- report- guidelines. *Association for Psychological Science*. https: //www.psychologicalscience.org/publications/replication.

Arenburg, David (26. Dez. 2016). *R legend for color density scatterplot produced using smooth Scatter.* https:/ /stackoverflow.com/questions/14271584/r-legend - for - color-density-scatterplot-produced-using-smoothscatter.

Aslam, Naeem (4. Apr. 2014). *What is the acceptable range of skewness and kurtosis for normal distribution of data?* [Discussion]. url: https://www.researchgate.net/post/ What_is_the_acceptable_range_of_- skewness_and_kurtosis_for_normal_ distribution_of_data.

APA — American Psychological Association, (1996-12-15). I*nitial Report: Task Force on Statistical Inference*. Board of Scientific Affairs. Techn. Ber. https://www.apa.org/science/ leadership/bsa/statistical/tfsi-initial-report.pdf.

Azur, Melissa J. u. a. (2011). /Multiple Imputation by Chained Equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research 20*(1), 40–49. https://www.ncbi.nlm.nih.gov-/pmc/articles/PMC3074241/pdf/nihms267760.pdf. (Visited on 28. 06. 2019).

Bååth, Rasmus (26. Okt. 2014). *Probable Points and Credible Intervals, Part I: Graphical Summaries*. http :// www . sumsar . net / blog / 2014 / 10 / probable - points - and - credible-intervals-part-one (visited on 28. 05. 2019).

Bååth, Rasmus (8. Jan. 2015). *Probable Points and Credible Intervals, Part 2: Decision Theory*. http://www.sumsar.net/blog/2015/01/probable- points- and- credible- intervals-part-two (visited on 28. 05. 2019).

Bacher, J. (1994). *Clusteranalyse: Eine anwendungsorientierte Einführung*. München: Oldenbourg.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, 66(6), S. 423–437. http://stats.org.uk/statistical- inference/Bakan1966.pdf

Baker, Alan (2016). "Simplicity". *The Stanford Encyclopedia of Philosophy*. Hrsg. von Edward N. Zalta. Winter 2016. Metaphysics Research Lab, Stanford University.

Barstead, Matthew (7. Juni 2018). *Power Analyses for an Unconditional Growth Model using lmer*. url: https://www.deadreckoning.consulting/blog/2018/06/07/2018- 06- 07 - power - analyses - for - an - unconditional - growth - model - using - lmer/ (visited on 03. 07. 2020).

Bartlett, M.S. (1957). A comment on D.V. Lindley's statistical paradox. *Biometrika* 44, S. 533– 534.

Bartoš, František & Ulrich Schimmack (Nov. 2020). Z-curve 2.0: *Estimating replication rates and discovery rates*. https://replicationindex.files.wordpress.com/2020/11/ zcurve.2.0.v11.22.pdf (visited on 11/2020).

Bates, Douglas (20. Mai 2006). [R] *lmer, p-values and all that*. url: https : / / tolstoy . newcastle.edu.au/R-/help/06/05/27666.html.

Bateson, Gregory (1985). *Ökologie des Geistes. Anthropologische, psychologische, biologische und epistemologische Perspektiven.* Frankfurt am MaSuhrkamp.

Bauer, Axel W. (2000). *Deduktion, Induktion, Abduktion und die hypothetisch-deduktive Methode in den empirischen Wissenschaften*. http : / / www . uni - heidelberg . de / institute/fak5/igm/g47/bauerabd.htm (visited on 26. 06. 2001).

Baumgartner, Michael & Rüdi Epple (2013). A coincidence analysis of a causal chain: The Swiss Minaret vote. *Sociological Methods and Research* 43(2), S. 280–312.

Baumgartner, Michael & Alrik Thiem (2015). "Identifying Complex Causal Depenncies in Configurational Data with Coincidence Analysis". *The R Journal* 7(1), S. 76–184. https : / / journal .r- project . org / archive / 2015 / RJ - 2015 - 014 / RJ - 2015 - 014 . pdf

Bayes, Rev. Thomas (1763). "An Essay toward solving a Problem in the Doctrine of Chance [pu- blished posthum]". *Philosophical Transactions of the Royal Society of London* 53, S. 370–418. https : //www.york.ac.uk/depts/maths/histstat/essay.pdf,

Beaudouin, V. (2003). Statistical analysis of textual data: Benzécri and the French school of data analysis. *Glottometrics 33*, 56–72.

Beck-Bornholdt, Hans-Peter & Hans-Hermann Dubben (2003). *Der Schein der Weisen. Irrtümer und Fehlurteile im täglichen Denken*. Rowohlt Tb. ISBN: 9783499614507.

Behrens, W.V. (1929). Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen. *Landwirtschaftliche Jahrbücher 68*, 807–837.

Bellhouse, D.R: (2004). The Reverend Thomas Bayes, FRS: A Biography to Celebrate the Tercentenary of His Birth. *Statistical Science 19*(1), 3–43. https://www.york.ac.uk/depts/ maths/histstat/bayesbiog.pdf

Bem, Daryl J. (2011a). Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect. *Journal of Personality and Social Psychology 100*(3), 407–425. http : / / dx . doi . org / 10 . 1037 / a0021524 , %202019 - 06 - 16 : %20 % 5Curl % 7B https : / / pdfs . semanticscholar . org / 79ec / e4f787af713d82924e41d8c17ab130f4b22d.pdf%7D.

Bem, Daryl J. (2011b). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology 100*(3), 407–425. ISSN: 0022-3514. doi: 10.1037/a0021524. url: http://dx.doi.org/10.1037/a0021524.

Bem, Daryl J., Jessica Utts & Wesley O. Johnson (2011). Reply — Must Psychologists Change the Way They Analyze Their Data? *Journal of Personality and Social Psychology, 101*(4), 716–719. http://www.deanradin.com/evidence-/Bem2011.pdf

Benjamin, Daniel J. et al. (2018-01-01). Redefine statistical significance. We propose to change the default P-value threshold for statistical significance from 0.05 to 0.005 for claims of new dis- coveries. *Nature Human Behaviour 2*(1), S. 6–10. https :// www . nature . com / articles /s41562-017-0189-z.pdf.

Benjamin, Lorna S. (1974). Structural analysis of social behavior. *Psychological Review, 81*, 392–425.

Benjamin, Lorna S. (1993). *Interpersonal Diagnosis and Treatment of Personality Disorder*. New York: Guilford Press.

Bennett, Craig M. et al. (2009). Neural Correlates of Interspecies Perspective Taking in the Post- Mortem Atlantic Salmon: An Argument For Multiple Comparisons Correction. *submitted*. http://prefrontal.org/files/posters/Bennett-Salmon--2009.pdf.

Bennett, Max R. & Peter M. S. Hacker (2003). *Philosophical Foundations of Neuroscience*. Oxford: Blackwell Publishers.

Benzécri, J.-P. (2012). *Correspondence analysis handbook*. New York: Marcel Dekker, Inc.

Berelson, Bernard (1952). *Content analysis in communications research*. In , James O. Berger (Ed.) (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer.

Berger, James O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis* 1(3)S. 385–402. https://projecteuclid.org/euclid.ba/1340371035

Berger, James O. (2014). Objective Bayesian Analysis. Lecture at the conference »*Priors, Quaternions, und Residuals, Oh, My!*« — *a scientific meeting in honor of retiring professor William H. Jefferys*. University of Texas at Austin. 2014-09-24. http://www.as.utexas.edu/jefferys/slides/berger.pdf.

Bergman, Manfred Max & Anthony P.M. Coxon (2005). The Quality in Qualitative Methods [54 paragraphs]. Forum *Qualitative Sozialforschung/ Forum Qualitative Social Research* [Online journal] 6(2) Art.34. http : / / nbn - resolving . de / urn : nbn : de : 0114 -fqs0502344.

Bermeitinger, Christina u. a. (2016). Positionspapier zur Lage der Allgemeinen Psychologie. *Psychologische Rundschau 67*(3), 175–179.

Bernoulli, Jakob (1713). *Ars conjectandi* [wieder abgedruckt Die Werke von Jakob Bernoulli (1975). Vol. 3. Birkhäuser: Basel, S.I0J–286]. Basel: Thurnisiorum.

Berry, Andrew C. (1941). The Accuracy of the Gaussian Approximation to the Sum of Independent Variates. *Transactions of the American Mathematical Society* 49(1), S. 122–136.

Betancourt, Michael (2017). *A Conceptual Introduction to Hamiltonian Monte Carlo* [latest version: 20I8-0J-I6]. https :// arxiv . org / abs / 1701 . 02434. University of Warwick, Coventry CV4 7AL, UK: Centre for Research in Statistical Methodology, Applied Statistics Center at Columbia University. (Visited on 05. 06. 2019).

Betancourt, Michael (2017-01). *How the Shape of a Weakly Informative Prior Affects Inferences*. https : / / mc - stan . org / users / documentation / case - studies / weakly _ informative_shapes.html. (Visited on 23. 07. 2019).

Bicare, Ned (21. Juni 2016). *Introducing the p-hacker app: Train your expert p-hacking skills*. P-hacker app: http://shinyapps.org/apps/p-hacker. url: https://www.nicebread.de/ introducing-p-hacker (visited on 22. 06. 2019).

Bickel, J., Hammel. E.A. & J.W. O'Connell (1975). Sex bias in graduate admissions: Data from Berkeley. *Science 187* (4175), 398–404. http://brenocon.com/science_1975_sex_bias_ graduate_admissions_data_berkeley.pdf

Bickel, P.J. & B.J.K. Kleijn (2012). The semiparametric Bernstein-von Mises Theorem. *The Annals of Statistics, 40*(1), 206–237.

Birnbaum, Allan (1977). The Neyman-Pearson Theory as Decision Theory, and as Inference Theo- ry; With a Criticism of the Lindley-Savage Argument for Bayesian Theory. *Synthese* 36(1), S. 19–49. https://www.pp.rhul.ac.uk-/ ~ cowan/stat/pcl/Birnbaum1977.pdf

Bittner-Stephan, Anna (2015). *Vergleich der Forschung zweier wissenschaftlicher Journals nach Kriterien der Replizierbarkeit* [Bachelorarbeit]. Techn. Ber. https://osf.io/y8tb3. München: Ludwig-Maximilians-Universität München, Department Psychologie, Lehrstuhl Allgemeine Psychologie IIMU.

Bland, J. Martin & Douglas G. Altman (1986). Statistical methods for assessing agreement bet- ween two methods of clinical measurement. Lancet i. http : / / www - users . york . ac.uk/ ~ mb55/meas/ba.pdf, reprinted with slight corrections in *Biochimica Clinica*, 11, 399–404, S. 307–310. (Visited on 28. 06. 2019).

Blei, David M. (2011-12-16). *Posterior Predictive Checks*. COS597C: Advanced Methods in Probabilistic Modeling. Fall, 2011. https :// www . cs . princeton . edu / courses / archive / fall11/cos597C/lectures/ppc.pdf. Princeton University.

Bock, Hans Hermann (1974). *Automatische Klassifikation. Theoretische und praktische Methoden zur Gruppierung und Strukturierung von Daten (Cluster-Analyse)*. Göttingen: Vandenhoeck und Ruprecht.

Bode, Sabine (2005). *Die vergessene Generation. Die Kriegskinder brechen ihr Schweigen*. München: Piper.

Bode, Sabine (2008). *Die deutsche Krankheit — German Angst*. München: Piper.

Boehmke, Bradley (2019). *Hierarchicaland K-means Cluster Analysis*[UC Business Analytics RPro gramming Guide]. https://uc-r.github.io/hc_clustering, https://uc-r.github.io/kmeans_clustering. (Visited on 20. 05. 2019).

Bohannon, John (2015). Many psychology papers fail replication test. *Science* 349(6251), S. 910– 911.

Bohm, David (1998). *Der Dialog. Das offene Gespräch am Ende der Diskussion*. Stuttgart: Klett- Cotta.

Bolker, Ben u. a. (5. Mai 2019). *GLMM FAQ*. url: http : / / bbolker . github . io /mixedmodels-misc/glmmFAQ.html.

Bolker, Ben et al. (9. Jan. 2020). *GLMM FAQ*. url: https : / / bbolker . github . io / mixedmodels--misc/glmmFAQ.html#testing-hypotheses (visited on 15. 10. 2020).

Bollen, Kenneth A. u. a. (2011). A Comparison of Bayes Factor Approximation Methods Including Two New Methods. *Sociological Methods and Research* 41(2), S. 294–324. http :// math.bu.edu/people/sray/preprints/smr_MS242_-mar10.pdf.

Bolstad, William M. (2007). *Introduction to Bayesians Statistics*. Hoboken, New Jersey: John Wiley & Sons, Inc.
Bonett, D.G. & E. Seier (2003). Statistical inference for a ratio of dispersions using paired samples. *Journal of Educational and Behavioral Statistics* 28, S. 21–30.
Bortz, Jürgen (1993). *Statistik: Für Sozialwissenschaftler* (Springer-Lehrbuch) (German Edition). Springer. ISBN: 9783540562009.
Box, George E.P. (1980). Sampling and Bayes' Inference in Scientific Modelling and Robustness. *Journal of the Royal Statistical Society. Series A (General) 143*(4), 383–430. https : / / www . cs . princeton . edu / courses / archive / fall11 / cos597C / reading / Box1980 . pdf
Boyd, Robert & Peter J. Richerson (2009). Culture and the evolution of human cooperation. *Philosophical Transactions of the Royal Society. Series B 364*(1533), 3281–3288. https://www.ncbi.nlm. nih.gov/pmc/articles/PMC2781880-/pdf/rstb20090134.pdf
Bozdogan, Hamparsum (2000). Akaike's Information Criterion and Recent Developments in In- formation Complexity. Journal of Mathematical Psychology, *44*, 62–91. http://yaroslavvb. com/papers/bozdogan-akaike.pdf
Brandstätter, Eduard (1999). Konfidenzintervcalle als Alternative zu Signifikanztests. *Methods of Psychological Research Online 4*(2). https://www.dgps.de/fachgruppen/methoden/ mpr-online/issue7/art3/brandstaetter.pdf.
Breiman, L. (2001). Statistical Modeling: the two cultures. *Statistical Science 16*(3), 199–231. Data set on BUPA liver disorders created by BUPA Medical Research Ltd., donored by Richard S. Forsyth, 1990-05-15
Breiman, L. et al. (1993). *Classification and regression trees*. New York: Chapman und Hall.
Bretthorst, G. Larry (1993). On the difference in means. In W.T. Grandy & P.W. Milonni (Eds..), *Physics and Probability Essays in honor of Edwin T. Jaynes*, pp. 177–194. http://bayes.wustl.edu/ glb/diff.pdf. Cambridge: Cambridge University Press
Breusch, T.S. & A.R. Pagan (1979). A simple test for heteroscedasticity and random coefficient variation. *Journal of the Econometric Society, Econometrica 7*(4), 1287–1294.
Brooks, Stephen P. & Andrew Gelman (1998). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics, 7*(4), 434–455. http:// www2.stat.duke.edu/ ~ scs/Courses/-Stat376/Papers/ConvergeDiagnostics/ BrooksGelman.pdf
Brown, Morton B & Alan B. Forsythe (1974). Robust tests for equality of variances. *Journal of the American Statistical Association 69*, S. 364–367.
Browne, William J., Mousa Golalizadeh Lahi & Richard M.A. Parker (März 2009). *A Guide to Sample Size Calculations for Random Effect Models via Simulation and the MLPowSim Software Package*. Executable: http://www.bristol.ac.uk/cmm/software/mlpowsim, user guide: http://www.bristol.ac.uk/media--library/sites/cmm/migrated/ documents/mlpowsim-manual.pdf. (Visited on 22. 06. 2019).
Brunner, Jerry & Ulrich Schimmack (2018a). Estimating population mean power under condi tions of heterogeneity and selection for significance. *Meta-Psychology submitted for publication*. Submitted manuscript: https://replicationindex.files.wordpress.com/2018/10/final- revision- 874- manuscript- in- pdf- 2236- 1- 4- 20180425-mva-final-002.pdf.
Brunner, Jerry & Ulrich Schimmack (18. Okt. 2018b). *Latest R-Code to run Z-Curve* [updated 18/11/17, 35 lines of code]. https://replicationindex.files.wordpress. com/2018/10/z-curve-public-18-10-2811.docx.
Bryman, A. (2008). The end of the paradigm wars? In P. Alasuutari, L. Bickman & J. J. Brannen.(Eds..), The SAGE *Handbook of social research methods*, pp. 13–25. London: Sage
Brysbaert, Marc & Michaël Stevens (2018). Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of Cognition* 1(1): 9, 1–20. https://www.journalofcognition. org/articles/10.5334/joc.10 (Visited on 22. 06. 2019).
Buchner, Axel, Edgar Erdfelder & Franz Faul (1996). Teststärkeanalysen. In Edgar Erdfelder et al. (Eds..), Handbuch Quantitative Methoden (pp. 123–136).einheim: Beltz: PVU. Kap. I,.
Bulletin, The IMS (1988). Thomas Bayes, F.R.S. — 1701?–1761. Who Is this gentleman? When and where was he born? *The IMS Bulletin, 17(3), S. 4276–278*. https://www.york.ac.uk/depts/maths/histstat/ bayespic.htm
Bürkner, Paul (2019-05-23). *brms: Bayesian Regression Models using 'Stan'* [version 2.9.0]. Vignettes from the R package brms. https://cran.r- project.org/web/packages/ brms/index.html.
Buuren, Stef van (2018). *Flexible Imputation of Missing Data*. Second Edition. Boca Raton/ FL: Chapman & Hall/CRC.
Buuren, Stef van & Karin Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software 45*(3), 1–67. https : / / www . jstatsoft . org / index . php / jss / article / view / v045i03 / v45i03 . pdf
Camerer, Colin u. a. (Aug. 2018). Evaluating the replicability of social science experiments in Nature and Science between 20I0 and 20I5. SocArXiv 4hmb6. Published 2018 in *Nature Human Behaviour, 2, 637–644*, https://www.nature.com/articles/s41562-018-0399-z/. Center for Open Science. doi: 10.31219/osf.io/4hmb6. url: https://ideas.repec. org/p/osf/socarx/4hmb6.html.
Camic, Paul M., Jean E. Rhodes & Lucy Yardley (2003). Naming the stars: Integrating qualitative methods into psychological research. In P. M. Camic, J. E. Rhodes, & L. Yardley (Eds.), *Qualitative research in psychology: Expanding perspectives in methodology and design* (pp. 3–15). Washington, American Psychological Association.
Campbell, D. & D. Fiske (1959). Convergant and discriminant validation by multitrait- multimethod matrix. *Psychological Bulletin 56*, S. 81–105.
Carnap, Rudolf (1952). *The Continuum of Inductive Methods*. Chicago: University of Chicago Press.

Carnap, Rudolf (1973). *Grundlagen der Logik und Mathematik*. München: Nymphenburger Verlagshandlung.
Carnap, Rudolf & Wolfgang Stegmüller (1959). *Induktive Logik und Wahrscheinlichkeit*. Wien: Nymphenburger Verlagshandlung.
Carroll, Harriet A. et al.(2017). The perceived feasibility of methods to reduce publication bias. *PLoS One 12*(10). https://journals.plos.org/plosone/article?id=10.1371/ journal.pone.0186472.
Casella, George (1985). An Introduction to Empirical Bayes Data Analysis. *The American Statistician 39*(2), 83–87. url: http : / / www . biostat . jhsph . edu / ~fdominic / teaching/bio656/labs/labs09/Casella.EmpBayes.pdf.
Casella, George (1992). Explaining the Gibbs Samples. *The American Statistician* 46(3). http://biostat.jhsph.edu/-~mmccall/articles/casella_1992.pdf, S. 167–174.
Caticha, Ariel (2009). *Quantifying Rational Belief* [Presented at MaxEnt 2009, the 29th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (July 5–10, 2009, Oxford, Mississippi, USA)]. https://arxiv.org/pdf/0908.3212.pdf.
Cecchin, Gianfranco, Gerry Lane & Wendel A. Ray (2002). *Respektlosigkeit*. Heidelberg: Carl-Auer Systeme.
Chakravartty, Anjan (2016). *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition). In Edward N. Zalta (Hrsg.). Chap. Scientific realism. url: https://plato.stanford. edu/archives/win2016/%20entries/scientific-realism.
Charmaz, Kathy (2006). *Constructing Grounded Theory. A Practical Guide Through Qualitative Analysis*. London: Sage.
Charmaz, Kathy (2012). *Constructing Grounded Theory*. London: Sage.
Chen, Yong & Sheng Luo (2011). A few remarks on 'Statistical distribution of the difference of two proportions'. *Statistics in Medicine 30*(15), 1913–1915.
Chick, Stephen (2005). Subjective Probability and Bayesian Methodology. In Shane G. Henderson & Barry L. Nelson (Eds.) *Handbook in Operations Research and Management Science: Simulation Vol. I3.*(pp. 225-257). Elsevier, S. url: https://epdf.tips/handbooks-in-operations-research-and-management- science - volume - 13 - simulation . html.
Chivers, Corey (6. Feb. 2012a). *General Bayesian estimation using MHadaptive*. url: https:// bayesianbiologist.com/-2012/02/06/general-bayesian-estimation-using- mhadaptive (visited on 05. 06. 2019).
Chivers, Corey (10. Feb. 2012b). *Visualising the Metropolis-Hastings algorithm*. url: https://bayesianbiologist . com / 2012 /02/10/visualising-the-metropolis-hastings-algorithm-2 (visited on 05. 06. 2019).
Christensen, Ronald (2010-07-16). *Revised Lindley-Jeffreys Paradox: Section 4.I.3*. https://www. ics.uci.edu/~wjohnson/-BIDA/Ch4/Lindley-paradox.pdf. Department of Mathematics u. a.
Clark, Michael (2016). *MCMC Algorithms* [latest version 20I6-II-2J]. Techn. Ber. https ://m- clark . github . io / docs / ld _mcmc. Advanced Research Computing consulting group, CSCAR, University of Michigan.
Clark, Michael (2018). *Bayesian Basics* [latest version: 20I8-0J-30]. Techn. Ber. https://m-clark. github . io / bayesian - basics. Advanced Research Computing consulting group, CS-CAR, University of Michigan.
Clyde, Merlise et al. (2019-03-29). *An Introduction to Bayesian Thinking. A Companion to the Sta- tistics with R Course*. Techn. Ber. R-code: https : / / github . com / StatsWithR / book. Coursera Inc. online learning platform [Course: Master Statistics with R. Statistical mastery of data analysis including inference, modeling, and Bayesian approaches. url: https : / / statswithr.github.io/book (visited on 28. 05. 2019).
Coghlan, Avril (2017). *A Little Book of R for Bayesian Statistics* [Release 0.I, latest version: 20IJ-II-0J]. Techn. Ber. Welcome Trust Sanger Institute, Cambridge/ UK. url: https://buildmedia. readthedocs . org / media / pdf / a - little - book - of - r - for - bayesian - statistics/latest/a- little- book- of- r- for- bayesian- statistics.pdf (visited on 04. 05. 2019).
Cohen, Jacob (1962). The statistical power of abnormal-social psychological research: a review. *Journal of Abnormal and Social Psychology 65*(3), S. 145–153.
Cohen, Jacob (1969). *Statistical Power Analysis for the Behavioral Sciences*, Ist ed. (2nd ed. 1988). Hillsdale: Lawrence Erlbaum Associates.
Cohen, Jacob (1992). A Power Primer. *Psychological Bulletin 112*(1), S. 155–159.
Cohen, Jacob (1994). The earth is round ($p < .05$). *American Psychologist 49*(12), 997–1003.
Collaboration, Open Science (2015). Estimating the reproducibility of psychological science. *Science 349*(6251). Author post-print: https://curve.coventry.ac.uk/open/file/2cb6ae15-530f-49eb-9b32-87ea1ee9493c-/1/Psycho-logical%20science.pdf, supplemental material https://science.sciencemag.org/content/sci/suppl/2015/08/26-ö/349.6251.aac4716.DC1/Aarts-SM.pdf.
Concar, David (2014). An investigation of the false discovery rate and the misinterpretation of p- values. *Royal Society Open Science 1*(3). url: https :// royalsocietypublishing . org/doi/pdf/10.1098/rsos.140216 (visited on 28. 05. 2019).
Concar, David (2017). The reproducibility of research and the misinterpretation of p-values. *Royal Society Open Science 4*(12). url: https://royalsocietypublishing.org/doi/ pdf/10.1098/rsos.171085 (visited on 28. 05. 2019).
Concar, David (2018). The False Positive Risk: A proposal concerning what to do about p-values [last revised 2019-01-10 v6]. *The American Statistician. Issue supl: Statistical Inference in the 2Ist Century: A World Beyond $p < 0.05$ 73*, 192–201. url: https://arxiv.org/pdf/1802.04888.pdf (visited on 28. 05. 2019).
Conover, W.J. & R.L. Imam (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician 35*, 124–129.
Coppock, Alexander, Jasper Cooper & Neal Fultz (2019-04-23). *randomizr: Easy-to-Use Tools for Common Forms of Random Assignment and Sampling* [version 0.I8.0] [Vignette 'Design and Analysis of Experiments with randomizr']. Techn. Ber. Vignette: https :// cran .r-project . org / web / packages / randomizr / vignettes / randomizr _ vignette . html. CRAN. url: https :// cran .r- project . org / web / packages / randomizr (visited on 27. 06. 2019).

countrymeters (2019). *countrymeters — Deutschland Bevölkerung*. url: https : / / countrymeters.info/de/Germany (visited on 22. 05. 2019).

Cox, Richard T. (1946). Probability, Frequency and Reasonable Expectation. *American Journal of Physics*, *14*(1). https://wwwusers.ts.infn.it/ ~ milotti/Didattica/Bayes/ Cox_1946.pdf, S. 1–13.

Cox, Richard T., Hrsg. (1961). *The Algebra of Probably Inference*. Baltimore: John Hopkins Press.

Creswell, John W (2009). *Research design: Qualitative, quantitative, and mixed methods approaches*. 3rd ed. Los Angeles: Sage.

Creswell, John W. & Vicki L. Plano Lark (2011). *Designing and Conducting Mixed Methods Research*. Chapter 3: https :/ / www . sagepub . com / sites / default / files / upm - binaries/35066._Chapter3.pdf. Thousand Oaks: Sage. (Visited on 04. 06. 2019).

Cronqvist, Lasse (2019). *Tosmana* [Version I.6I]. https :// www . tosmana . net. University of Trier. Trier.

Dablander, Fabian (3. Aug. 2015). *Bayesian Statistics: Why and How*. url: https : / / blog . efpsa . org / 2015 / 08 / 03 / bayesian - statistics - why - and - how (visited on 04. 06. 2019).

Dalgaard, Peter (2004). *Introductory Statistics with R* (Statistics and Computing). Springer. isbn: 9780387954752.

Darwin, Charles R. (1876). *The effects of cross and self fertilisation in the vegetable kingdom*. http://darwin-online.org.uk/converted/published/1881_Worms_F1357/1876_ Cross and Self Fertilisation _ F1249 / 1876 _ CrossandSelfFertilisation _ F1249.html. London: John Murray.

Das, Abhranil (8. Feb. 2014). *R Code for multivariate random-walk Metropolis sampling.* url: https : / / blog . abhranil .net/2014/02/08 / r - code - for - multivariate - random-walk-metropolis-hastings-sampling (visited on 05. 06. 2019).

Dave Kincaid https://stats.stackexchange.com/users/118/dave- kincaid, user (22. Juni 2011). *Calculating the parameters of a Beta distribution using the mean and variance.* Cross Validated. url: https :/ / stats . stackexchange . com / questions / 12232 / calculating- the- parameters- of- a- beta- distribution- using- the- mean- and-variance (visited on 04. 06. 2019).

Davidson-Pilon, Cameron (4. Nov. 2015). *Bayesian Methods for Hackers: Would You Rather Lose an Armora Leg?* Chapter 5 from the book Bayesian Methods for Hackers (2015). informIT — the trusted technology learning source [Pearson Education]. url: http://www.informit.com/articles/article.aspx?p = 2447200&seqNum = 2.

Dawson, Robert J. MacG. (1995). The unusual episode Data Revisited. *Journal of Statistic Education 3*(3). http : / / jse . amstat . org / v3n3 / datasets . dawson . html. doi: 10.1080/10691898.1995.11910499.

Deary, Ian J., Alison Pattie & John M. Starr (2013). The Stability of Intelligence from Age 11 to Age 90 Years: The Lothian Birth Cohort of 1921. *Psychological Science 24*(12), S. 2361–2368.

Deary, Ian J., Lawrence J. Whalley u. a. (2000). The stability of individual differences in mental ability from childhood to old age: Follow-up of the 1932 Scottish Mental Survey. *Intelligence 28*, 49–55.

Deary, Ian J., Martha C. Whiteman u. a. (2004). The Impact of Childhood Intelligence on Later Life: Following Up the Scottish Mental Surveys of 1932 and 1947. *Journal of Personality and Social Psychology 86(1)*, 130–147. ISSN: 0022-3514. doi: 10.1037/0022-3514.86.1.130. url: http://dx.doi.org/10.1037/0022-3514.86.1.130.

deFinetti, Bruno (1974). *Theory of Probability Vol. I.* New York: John Wiley und Sons.

Dempster, A.P., N.M. Laird & Donald B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*(1), 1–38.

Dempwolff, Otto (1939). Grammatik der Jabêm-Spracheauf Neuguinea. Friederichsen: de Gruyter.

Derringer, George & Ronald Suich (1980). Simultaneous Optimization of Several Response Variables. *Journal of Quality Technology — A Quarterly Journal of Methods, Applications and Related Topics 12*(4), 214–219.

Deutschland, Bundesrepublik (2. Juli 2020). Bevölkerungsstand. url: https://www.destatis.de / DE / Themen / Gesellschaft - Umwelt / Bevoelkerung / Bevoelkerungsstand /_inhalt.html (visited on 02. 07. 2020).

Diaconis, Persi & David A. Freedman (1986). On the Consistency of Bayes Estimates. *The Annals of Statistics 14*(1), 1–26.

Dienes, Zoltan, Simon Coulton & Nick Heather (2018). Using Bayes factors to evaluate evidence for no effect: examples from the SIPS project. *Addiction 113*(2), 240–246. url: http://nrl.northumbria.ac.uk/32294/1/-Dienes%5C%20et%5-C%20al%5C%20%5C%20%5C%20accepted%5C%20version.pdf (visited on 23. 05. 2019).

Diepgen, Raphael (1999). Warum nur n-1 und nicht n? Erwartungstreue — leicht gemacht. *Stochastik in der Schule 19*(1), 10–13. https: / / www . stochastik - in - der - schule . de / sisonline/struktur/jahrgang19-99/heft1-/1999-1_Diepgen.pdf

Dixon, Philip M. u. a. (2018). A primer on the use of equivalence testing for evaluating measure ment agreement. *Med Sci Sports Exerc 50*(4), 837–845. url: https :// www.ncbi.nlm.nih.gov/pmc/articles/PMC5856600/pdf/-nihms916849.pdf.

Dobson, Annette J., Hrsg. (1990). *An introduction to generalized linear models*. https : / / reneues.files.wordpress.com/-2010/01/an-introduction-to-generalized- linear-models-second-edition-dobson.pdf. Boca Raton/ Florida: Chapman und Hall/ CRC.

Dolinski, Dariusz et al. (2017). Would You Deliver an Electric Shock in 2015? Obedience in the Experimental Paradigm Developed by Stanley Milgram in the 50 Years Following the Original Studies. *Social Psychological and Personality Science 8*, S. 927–933.

Doob, Joseph L. (1949). Application of the theory of martingales. *Colloq. Intern. du C.N.R.S (Paris) 13*, S. 23–27.

Duane, Simon et al. (1987). Hybrid Monte Carlo. *Physical Letters B 1985*, S. 216–222.

Durante, Kristina M., Ashley Rae Arsena & Vladas Griskevicius (2013). The fluctuating female vote: Politics, religion, and the ovulatory cycle. *Psychological Science 24,* 1007–1016. https://www. researchgate.net/publication/-236277144_The_Fluctuating_Female_Vote

Dus, Adrian (2017). *Consistency Cubes: A Fast, Efficient Method for Boolean Minimization* [Working Paper, Version 4, 20IJ-0J]. url: https://www.researchgate.net/publication/ 316081582 _ Consistency _ Cubes _A_ Fast _ Efficient _ Method _ for _ Boolean _ Minimization (visited on 20. 03. 2019).

Dziak, John J. u. a. (2012). Sensitivity and specificity of information criteria. *Technical Report Series* #12-119. Pre-print from 2019-01-07: Dziak, John J., Coffman, Donna L., Lanza, Stephanie T., Li, Runze & Jermiin, Lars S. https://www.researchgate.net/publication/328435982_Sensitivity_and_Specificity_of_Information_Criteria. College of Health und Human Development, The Pennsylvania State University: The Pennsylvania State University. url: https://www.methodology.psu.edu/files/2019/03/12- 119-2e90hc6.pdf (visited on 29. 05. 2019).

Eddelbüttel, Dirk (2012). *Simulting pi from R or C++ in about five lines*. url: http://gallery. rcpp.org/articles/simulating-pi/ (visited on 04. 05. 2019).

Efron, Bradley (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics 7*(1). https : / / project-euclid . org / download / pdf _ 1 / euclid . aos / 1176344552, S. 1–26.

Efron, Bradley & Robert J. Tibshirani (1993). *An introduction to the boostrap*. Boca Raton/ Florida: Chapman & Hall CRC.

Ehrenberg, Andrew S.C. (1982). *A Primer in Data Reduction: An Introductory Statistics Textbook*. Chichester: John Wiley & Sons Ltd.

Eid, Michael, Mario Gollwitzer & Manfred Schmitt (2010). *Statistik und Forschungsmethoden*. 3. korrigierte Auflage (20I3). Weinheim: Beltz.

Eklund, Anders, Mats Andersson et al. (2012). Does parametric fMRI analysis with SPM yield valid results? — An empirical study of 1484 rest datasets. *NeuroImage 61*(3), S. 565–578.

Eklund, Anders, Thomas E. Nichols & Hans Knutsson (2016a). Can parametric statistical methods be trusted for fMRI based group studies? *PNAS 113*(28), S. 7900–7905.

Eklund, Anders, Thomas E. Nichols & Hans Knutsson (2016b). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *PNAS 113*(28), S. 7900–7905.

Enk, Steven J. van (2014-08-28). *The Brandeis Dice Problem and Statistical Mechanics*. Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 48(A). Draft submitted on 2019-05-22: https://arxiv.org/pdf/1408.6803.pdf, S. 1–6.

Erdfelder, Edgar & Jürgen Bredenkamp (1994). Hypothesenprüfung. In Theo Hermann & W.H. Tack (Hrsg.), *Enzyklopädie der Psychologie. Themenbereich B: Methdologie und Methoden. Serie I: Forschungsmethoden der Psychologie. Band I: Methodologische Grundlagen der Psychologie*. S. 604–648. Göttingen: Hogrefe,.

Erdfelder, Edgar & Rolf Ullrich (2018). Zur Methodologie von Replikationsstudien. *Psychologische Rundschau 69*(1), S. 3–21.

Erler, Nicole S. (2019-06-06). *JointAI: Joint Analysis and Imputation of Incomplete Data* [version 0.5.2]. Vignettes from the R package JointAI. https://cran.r-project.org/web/ packages/JointAI/index.html.

Erzberger, Christian & Udo Kelle (2003). Making inferences in mixed methods: The rules of integration. In Abbas Tashakkori & Charles Teddlie (Hrsg.), Handbook of mixed methods in social and behavioral research, Kap. 16, S. 457–488. Thousand Oaks: Sage.

Esseen, Carl-Gustav (1942a). A moment inequality with an application to the central limit theorem. *Skand. Aktuarietidskr 39*, S. 160–170.

Esseen, Carl-Gustav (1942b). On the Liapunoff limit of error in the theory of probability. *Arkiv för Matematik, Astronomi och Fysik A28*, S. 1–19.

Etz, Alexander (21. Mai 2015a). *Type-S and Type-M errors*. url: https://alexanderetz.com/ 2015/05/21/type-s--and-type-m-errors/ (visited on 29. 07. 2020).

Etz, Alexander (25. Juli 2015b). *Understanding Bayes: Updating priors via the likelihood*. url: https://alexanderetz.com/-2015/07/25/understanding- bayes- updating- priors-via-the-likelihood (visited on 04. 06. 2015).

Etz Alexander & Wagenmakers, Eric-Jan (2017). J.B.S. Haldane's Contribution to the Bayes Factor Hypothesis Test. *Statistical Science 32*(2), S. 313–329. https : / / projecteuclid . org / download/pdfview_1/euclid.ss/1494489818

Falbel, Daniel u. a. (5. Apr. 2019). *Tutorial: Overfitting and Underfitting*. keras: R Interface to 'Keras' [The Python Deep Learning library] [version 2.2.4.1]. url: https : / / cran . r -project.org/web/packages/keras/-vignettes/tutorial_overfit-_underfit. html (visited on 29. 05. 2019).

Faraway, Julian J. (2002-07). *Practical Regression and Anova using R*. https : / / cran . r - project.org/doc/contrib/Faraway--PRA.pdf. (Visited on 14. 10. 2020).

Faul, Franz, Edgar Erfelder, Axel Buchner et al. (2009). Statistical power analyses using GPower 3.1: Tests for correlation and regression analyses. *Behavioral Research Methods 41*(4), 1149–1160. http ://www.gpower.hhu.de/fileadmin/-redaktion/Fakultaeten/Mathematisch- Naturwissenschaftliche _ Fakultaet / Psychologie / AAP / gpower / GPower31 - BRM-Paper.pdf.

Faul, Franz, Edgar Erfelder, Albert-Georg Lang et al. (2007). GPower 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavioral Research Methods 39*(2), S. 175–191. https : / / link . springer . com / content / pdf / 10 . 3758 %2FBF03193146.pdf

Fay, Michael P. (2010). Confidence intervals that match Fisher's exact or Blaker's exact tests. *Biostatistics 11*(2), S. 373–374. https://academic.oup.com/biostatistics/article-pdf/11/2/373/18603640/kxp050.pdf.

Feng, Changyong et al. (2014). Log-transformation and its implications for data analysis. *Shanghai Archives of Psychiatry 26*(2), S. 105–109. https : / / www . ncbi . nlm . nih . gov / pmc / articles/PMC4120293/pdf/sap-26-02-105.pdf

Feyerabend, Paul (1976). Wider den Methodenzwang. Skizze einer anarchistischen Erkenntnistheorie. Frankfurt a. M.: Suhrkamp.

Fidler, Fiona (2010). The American Psychological Association Publication Manual Sixth Edition: Implications for Statistics Education. Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 20I0), Ljubljana, Slovenia. Hrsg. von C. Reading. https : / / iase -web . org / documents / papers / icots8 / ICOTS8 _ C156 _ FIDLER . pdf. Voorburg/ NL: International Statistical Institute.

Fiedler, Klaus (2018). Wo sind die wissenschaftlichen Standards für hochwertige Replikationsforschung? *Psychologische Rundschau 69*(1), S. 45–56.

Fienberg, Stephen E. (2006). When Did Bayesian Inference Become Bayesian? *Bayesian Analysis 1*(1), S. 1–40. https : / / projecteuclid . org / download / pdf _ 1 / euclid . ba / 1340371071

Fischler, Martin A. & Robert C. Bolles (1980). Random Sample Consensus: *A paradigm for model fitting with applications to image analysis and automated cartography* [Technial Note 2I3]. Techn. Ber. https://apps.dtic.mil-/dtic/tr/fulltext/u2/a460585.pdf, published 1981-06 in Comm. ACM, 24(6), p.381–395. Menlo Park, California, 94025, USA: Artificial Intelligence Center, SRI International. (Visited on 14. 10. 2020).

Fisher, Ronald Aylmer (1915). Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika 10*(4), S. 507–521.

Fisher, Ronald Aylmer (1921). On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron 1*, S. 3–32. https://digital.library.adelaide.edu.au/dspace/ bitstream/2440/15169/1/14.pdf

Fisher, Ronald Aylmer (1922). *On the Mathematical Foundations of Theoretical Statistics*. In: Philosophical Transactions Royal Society London. Series A 222A. , S. 309–368.

Fisher, Ronald Aylmer (1935). The fiducial argument in statistical inference. *Annals of Eugenics (London) 8*, S. 391–398. https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1469-1809.1935.tb02120.

Fisher, Ronald Aylmer (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics (London) 7*, S. 179–188. http : / / rcs . chemometrics . ru / Tutorials / classification/Fisher.pdf

Fisher, Ronald Aylmer (1955). Statistical Methods and Scientific Induction. *Journal of the Royal Statistical Society. Series B 17*, S. 69–78. https://www.phil.vt.edu/dmayo/personal_ website/Fisher-1955.pdf

Fisher, Ronald Aylmer (1956/1973). *Statistical Methods and Scientific Inference*. 3th ed. 1993. Published in »Statistical Methods, Experimental Design, and Scientific Inference. A Re-issue.« Edited by J.J Bennett in 1990, Oxford: Oxford Science Publications.

Fisher, Ronald Aylmer (1925/1973). *Statistical Methods for Research Workers*. I4th ed. 1993. Published in »Statistical Methods, Experimental Design, and Scientific Inference. A Re-issue.« Edited by J.J Bennett in 1990, reprinted 199I. Oxford: Oxford Science Publications.

Fisher, Ronald Aylmer (1935/1973). *The Design of Experiments*. 8th ed. 1966, reprinted by arrange- ment 199I. Published in »Statistical Methods, Experimental Design, and Scientific Inference. A Re-issue.« Edited by J.J Bennett in 1990, reprinted 199I. Oxford: Oxford Science Publications.

FiveThirtyEight (2019). *Hack Your Way To Scientific Glory*. url: https : / / projects .fivethirtyeight.com/p-hacking (visited on 22. 06. 2019).

Flanders, Ned A (1970). *Analyzing teacher behavior*. Reading (Mass.), Addison-Wesley PC.

Flick, Uwe (2000). Triangulation in der qualitativen Forschung. *Qualitative Forschung. Ein Handbuch*. Hrsg. von Uwe Flick, Ernst v. Kardorff & Ines Steinke. Reinbek bei Hamburg: Rowohlts Enzyklopädie, S. 309–318.

Flick, Uwe (2009). *An introduction to qualitative research* (4th ed.) London: Sage.

Flick, Uwe, Ernst v. Kardorff & Ines Steinke, Hrsg. (2000). *Qualitative Forschung. Ein Handbuch.* Reinbek bei Hamburg: Rowohlts Enzyklopädie.

Fowler, Dave (2019). *Titanic Facts. The Life and Loss of the RMS Titanic in Numbers* [A History in Numbers website: https://historyinnumbers.com]. url: https :// titanicfacts . net / life-on-the-titanic (visited on 20. 05. 2019).

Fox, John (2002). *An R and S-PLUS companion to applied regression*. http : / / socserv . mcmaster.ca/jfox/. Thousands Oaks, California: Sage Publications.

Fox, John & Sanford Weisberg (21. Sep. 2018). *Bootstrapping Regression Models in R*. An Appendix to An R Companion to Applied Regression, third edition [latest rev. 20I8-0!)-2I]. url: https://socialsciences.mcmaster.ca/jfox/Books/-Companion/appendices/Appendix-Bootstrapping.pdf (visited on 29. 07. 2020).

Fox, John & Sanford Weisberg (2019). *On-Line Appendices to An R Companion to Applied Re- gression*, third edition. Topics: Bayesian estimation of regression models, Bootstrapping regression models, Cox regression for survival data, Fitting regression models to survey data, Multiple im- putation of missing data, Multivariate linear models, Nonlinear regression, Nonparametric regression, Robust regression, Time-series regression [last modified: 20I8-0!)-28]. url: https ://socialsciences.mcmaster.ca/jfox/Books/Companion/appendices.html (be- sucht am 30. 07. 2020).

Francis, Gregory (2012). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin and Review 19*(2), S. 151–156. Pre-print: http : / /www2.psych.purdue.edu/ ~ gfrancis/- Publications/GFrancis- R1.pdf

Freedman, David A. (1963). On the Asymptotic Behavior of Bayes Estimates in the Discrete Case I. *The Annals of Mathematical Statistics 34*(4). https :// projecteuclid . org / download/pdf_1/euclid.aoms/1177703871, S. 1386–1403.

Freedman, David A. (1965). On the Asymptotic Behavior of Bayes Estimates in the Discrete Case II. *The Annals of Mathematical Statistics, 36*(2), 454–456.

French, Chris (15. März 2012). *Precognition studies and the curse of the failed replications*. url: https : / / www . theguardian . com / science / 2012 / mar / 15 / precognition - studies-curse-failed-replications (visited on 30. 07. 2020).

Fryar, C.D., Q. Gu & C.L. Ogden (2012). *Anthropometric reference data for children and adults: United States, 200J–20I0. Data from the National Health and Nutrition Examination Survey*. Techn. Ber. National Center for Health Statistics. Vital Health Statistics. url: https://www.cdc.gov/nchs/data/series/sr_11/sr11_252.pdf (visited on 06. 06. 2019).

Gabbatt, Adam (17. Jan. 2018). *A tall tale? Accuracy of Trump's medical report — and new height — questioned*. url: https://www.theguardian.com/us-news/2018/jan/17/a-tall- tale- accuracy- of- trumps- medical- report- and- new-height- questioned (visited on 17. 01. 2018).

Gabry, Jonah (2. Aug. 2018). *Graphical posterior predictive checks using the bayesplot package* [latest version: 2020-06-0I]. url: https://mc-stan.org/bayesplot/articles/graphical- ppcs.html (visited on 30. 07. 2020).

Gabry, Jonah & Ben Goodrich (2018). *How to Use the rstanarm Package* [version 2.I8.2] [latest ver- sion: 20I8-II-08]. rstanarm: Bayesian Applied Regression Modeling via Stan [Vignette]. https: //cran.r-project.org/web/packages/-rstanarm/vignettes/rstanarm.html# step-2-draw-from-the-posterior-distribution. CRAN.

Gabry, Jonah & Ben Goodrich (2018-04-13). *Prior Distributions for rstanarm Models*. rstanarm: Bayesian Applied Regression Modeling via Stan [Articles]. https : / / mc - stan . org / rstanarm/articles/priors.html. CRAN.

Gabry, Jonah, Daniel Simpson u. a. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 182*(2), S. 389–402. url: http://dx. doi.org/10.1111/rssa.12378.

Gaffert, Philipp, Florian Meinfelder & Volker Bosch (2016-01-25). *Towards an Multiple-Imputation-Proper Predictive Mean Matching*. Paper presented at Session 5I2: Various Flavors of Missing-Data Problems | Proceedings of the Joint Statistical Meetings (JSM) 20I8, Survey Research Methods Section, Vancouver, British Columbia, Canada, July 28 – August 2, 20I8. url: http://www.asasrms.org/Proceedings/y2018/files/867081.pdf (visited on 23. 06. 2019).

Gage, N. L. (1989). The paradigm wars and their aftermath. *Teachers College Record 91*(2), S. 135–150.

Gal, Shayanne & Samantha Lee (18. Feb. 2019). *The height differences between all the US presidents and first ladies*. Business Insider International. url: https://www.businessinsider. de / us - president - first - lady - height - differences - 2018 - 7?r = US & IR = T (visited on 06. 06. 2019).

Galak, Jeff, Robyn A. LeBoeuf et al. (2012-06-19). *Correcting the Past: Failures to Replicate Psi*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id = 2001721.

Galak, Jeff & Leif D. Nelson (2010-10-09). *A Replication of the Procedures from Bem (2010, Study 8) and a Failure to Replicate the Same Results*. SSRN. https://papers.ssrn. com/sol3/papers.cfm?abstract_id = 1699970.

Galavotti, Maria Carla (2001). *Subjectivism, Objectivism and Objectivity in Bruno de Finet- ti's Bayesianism. Foundations of Bayesianism*. Hrsg. von David Corfield & Jon Williamson. Dordrecht/ NL: Springer, S. 161–174. url: http : / / www . brunodefinetti.it / Bibliografia / Subjectivism , %5C % 20objectivism % 5C % 20and % 5C % 20objectivity % 5C % 20in % 5C % 20Bruno % 5C % 20de % 5C % 20Finetti % 5C % 20Bayesianism.pdf (visited on 07. 06. 2019).

Galton, Francis (1889). *Natural Inheritance*. http : / / galton . org / books / natural - inheritance/pdf/galton-nat-inh–1up-clean.pdf. London: Macmillan.

Galtung, Johan (1990). Theory formation in social research: A plea for pluralism. *Comparative methodology: Theory and practice in international social research*, 96–112.

Gelman, Andrew (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. International *Statistical Review 71*(2), 369–382. https://citeseerx.ist.psu.edu/ viewdoc / download ; jsessionid = 69EA431-0087D1B885BC556616A4387D3 ? doi = 10.1.1.6.9888&rep = rep1&type = pdf

Gelman, Andrew (29. Dez. 2004). *Type I, type 2, type S, and type M errors*. url: https://statmodeling.stat.columbia .edu/2004/12/29/type_1_type_2_t (visited on 24. 05. 2019).

Gelman, Andrew (16. Juni 2005). *Objective and Subjective Bayes*. url: https://statmodeling. stat.columbia.edu/2005/-06/16/objective_and_s (visited on 24. 05. 2019).

Gelman, Andrew (2007a). Comment: Bayesian Checking of the Second Levels of Hierarchical Models. *Statistical Science 22*(3), S. 349–352. http : / / www . stat . columbia . edu / ~ gelman / research/published/STS235A.pdf

Gelman, Andrew (31. Okt. 2007b). *Controversies over posterior predictive checks*. url: https://statmodeling.stat-.columbia.edu/2007/10/31/controversies_o (visited on 03. 06. 2019).

Gelman, Andrew (2008a). Objections to Bayesian statistics. *Bayesian Analysis 3*(3), S. 445–450. http ://www.stat.-columbia.edu/ ~ gelman/research/published/badbayesmain.pdf

Gelman, Andrew (12. Juni 2008b). *Some thoughts on the saying, All models are wrong, but some are useful*. url: https://statmodeling.stat.columbia.edu/2008/06/12/ all_models_are (visited on 24. 05. 2019).

Gelman, Andrew (7. Feb. 2009a). *Confusions about posterior predictive checks*. url: https : / / andrewgelman . com / 2009 / 02 / 07 / confusions _ abou (visited on 29. 05. 2019).
Gelman, Andrew (26. Feb. 2009b). *Why I don't like so-called Bayesian hypothesis testing*. url: https://statmodeling.stat.-columbia.edu/2009/02/26/why_i_dont_ like (visited on 28. 05. 2019).
Gelman, Andrew (2011a). *Induction and Deducation in Bayesian Data Analysis*. RMM[open access online journal]. Special Topic: Statistical Science and Philosophy of Science. Edited by Deborah G. Mayo, Aris Spanos and Kent W. Stale 2, S. 67–78. url: http : / / www . stat .columbia.edu/ ~ gelman/research/published/philosophy_online4.pdf.
Gelman, Andrew (2. Apr. 2011b). *So-called Bayesian hypothesis testing is just as bad as regular hypothesis testing*. url: https://statmodeling.stat.columbia.edu/2011/04/02/so- called_bayes (visited on 24. 05. 2019).
Gelman, Andrew (17. Mai 2013a). *How can statisticians help psychologists do their research better?* url: https : / / statmodeling . stat . columbia . edu / 2013 / 05 / 17 / how - can - statisticians-help-psychologists-do-their-research--better (visited on 24. 05. 2019).
Gelman, Andrew (11. März 2013b). *My problem with the Lindley paradox*. url: https : / / statmodeling . stat . columbia . edu / 2013 / 03 / 11 / my - problem - with - the - lindley-paradox (visited on 24. 05. 2019).
Gelman, Andrew (2013c). P-values and statistical practice [Commentary]. *Epidemiology 24*(1), S. 69–72. http://www.stat.columbia.edu/ ~ gelman/research/published/pvalues3. pdf
Gelman, Andrew (2013d). Two simple examples for understanding posterior p-values whose dis- tributions are far from unform. *Electronic Journal of Statistics 7*, S. 2595–2602. http://www.stat. columbia.edu/ ~ gelman/research-/published/ppc_understand3.pdf
Gelman, Andrew (15. März 2014a). *Problematic interpretations of confidence intervals*. url: https:/ / statmodeling . stat . columbia . edu / 2014 / 03 / 15 / problematic - interpretations-confidence-intervals (visited on 25. 05. 2019).
Gelman, Andrew (13. Feb. 2014b). *Stopping rules and Bayesian analysis*. url: https://statmodeling.stat.-columbia.edu/2014/02/13/stopping-rules-bayesian- analysis (visited on 25. 05. 2019).
Gelman, Andrew (11. Dez. 2014c). *The Fallacy of Placing Confidence in Confidence Intervals*. url: https://statmodeling.stat.columbia.edu/2014/12/11/fallacy-placing- confidence-confidence-intervals (visited on 25. 05. 2019).
Gelman, Andrew (17. Nov. 2014d). *This is what 'power = 0.06' looks like*. Get used to it. url: https://statmodeling.stat.-columbia.edu/2014/11/17/power- 06- looks- like- get-used (visited on 24. 05. 2019).
Gelman, Andrew (30. Juni 2014e). *Who invented the Metropolis algorithm?* url: https : / / statmodeling . stat . columbia . edu / 2014 / 06 / 30 / invented - metropolis - algorithm (visited on 05. 06. 2019).
Gelman, Andrew (27. Jan. 2015a). *About a zillion people pointed me to yesterday's xkcd cartoon*. url: https://statmodeling.stat.-columbia.edu/2015/01/29/six-quick-tips-improve-regression-modeling (visited on 29. 05. 2019).
Gelman, Andrew (12. Mai 2015b). *Happytalk, meetthe Edlinfactor* [Comment by Andrew Gelman]. url: https://statmodeling.stat.columbia.edu/2016/05/12/happy-talk- meet-the-edlin-factor/%5C#comment-272996 (visited on 03. 06. 2019).
Gelman, Andrew (29. Jan. 2015c). *Six quick tips to improve your regression modeling*. url: https://statmodeling.stat.-columbia.edu/2015/01/29/six-quick-tips-improve- regression-modeling (visited on 24. 05. 2019).
Gelman, Andrew (25. Mai 2015d). *The difference between significant and not significant is not itself statistically significant: Education edition*. url: https : / / statmodeling . stat . columbia . edu / 2016 / 05 / 25 / the - difference - between - significant - and -not-significant-is-not-itself-statistically-significant-education- edition (visited on 03. 06. 2019).
Gelman, Andrew (2017a). Honesty and transparency are not enough. *Chance 30*(1), S. 37–39. url: http : / / www . stat . columbia . edu / ~ gelman / research / published / ChanceEthics14.pdf.
Gelman, Andrew (4. Dez. 2017b). *The '80% power' lie*. url: https://statmodeling.stat. columbia.edu/2017/12/04/80--power-lie (visited on 24. 05. 2019).
Gelman, Andrew (3. März 2017c). *Yes, it makes sense to do design analysis ('power calculations') after the data have been collected*. url: https://statmodeling.stat.columbia.edu/2017/ 03 / 03 / yes - makes - sense - design - analysis - power - calculations - data - collected (visited on 30. 07. 2020).
Gelman, Andrew (24. Sep. 2018). *Don't calculate post-hoc power using observed estimate of effect size*. url: https : / / statmodeling . stat . columbia . edu / 2018 / 09 / 24 / dont - calculate-post-hoc-power-using-observed-estimate--effect-size (visited on 24. 05. 2019).
Gelman, Andrew (13. Jan. 2019a). *How posthoc power calculation is like a shit sandwich*. url: https://statmodeling.stat . columbia.edu/2019/01/13/post-hoc-power-calculation-like-shit-sandwich (visited on 24. 05. 2019).
Gelman, Andrew (2019b). *Statistical Modeling, Causal Inference, and Social Science*. url: https://statmodeling.stat.columbia.edu (visited on 24. 05. 2019).
Gelman, Andrew (27. Jan. 2019c). *What should JPSP have done with Bem's ESP paper, back in 20I0? Click to find the surprisingly simple answer!* url: https :// statmodeling . stat . columbia.edu/2019/01/27/jpsp-done-bem-paper-back--2010-click-find- surprisingly-simple-answer (visited on 24. 05. 2019).
Gelman, Andrew & John Carlin (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science 9*(6), S. 641–651. http://www. stat.columbia.edu/ ~ gelman/-research/published/retropower_final.pdf doi: 10.1177/1745691614551642.

Gelman, Andrew, John B. Carlin u. a. (2003). *Bayesian Data Analysis*, Second Edition (Chapman and Hall/CRC Texts in Statistical Science). Chapman & Hall/CRC. isbn: 158488388X. Gelman, Andrew, John B. Carlin u. a. (2004). Bayesian Data Analysis. Boca Raton, Florida: Chapman & Hall/CRC.

Gelman, Andrew, Ben Goodrich u. a. (2019). R-squared for Bayesian regression models. *The American Statistician 73*(3), S. 307–309. https : / / stat . columbia . edu / ~ gelman / research / published/bayes_R2_v3.pdf (Visited on 15. 10. 2020).

Gelman, Andrew & Jennifer Hill (2007). *Data Analysis Using Regression and Multilevel/ Hierar chical Models* (Analytical Methods for Social Research). Cambridge/ NJ: Cambridge University Press.

Gelman, Andrew, Jennifer Hill, Yu-Sung Su et al. (2015-06-16). *mi: Missing Data Imputation and Model Checking* [version 1.0]. Vignette from the R package mi. https : / / cran . r - project.org/web/packages/mi/index.html.

Gelman, Andrew, Jennifer Hill & Masanao Yajima (2012). Why We (Usually) Don't Have to Worry About Multiple Comparisons. *Journal of Research on Educational Effectiveness 5*(2), S. 189–211. doi: 10 . 1080 / 19345747 . 2011 . 618213. url: http :// www . stat .columbia.edu/ ~ gelman/research/published/multiple2f.pdf.

Gelman, Andrew, Jessica Hwang & Aki Vehtari (2013). Understanding predictive information criteria for Bayesian models. *Statistics and Computing 24*(6), S. 997–1016. Pre-print: http :// www . stat.columbia.edu/ ~ gelman/-research/published/waic_understand3.pdf

Gelman, Andrew & Eric Loken (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no fishing expedition or p-hacking and the research hypothesis was posited ahead of time* (published 20I3-II-I4). url: http://www.stat.columbia.edu/ ~ gelman/research/unpublished/p_hacking.pdf.

Gelman, Andrew & Eric Loken (2014). The statistical crisis in Science. Data-dependent analysis — a garden of forking paths – explains why many statistically significant comparisons don't hold up. *American Scientist 102*, S. 460–465. url: http :// www . stat . columbia . edu/ ~ gelman/research/published/ForkingPaths.pdf.

Gelman, Andrew, Xiao-Li Meng & Hal Stern (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica 6*, S. 733–807. http://www.stat.columbia. edu/ ~ gelman/research/published/-A6n41.pdf

Gelman, Andrew & Donald B. Rubin (1992). Inference from Iterative Simulation Using multiple sequences. *Statistical Science 7*(4), S. 457–472. http : / / www2 . stat . duke . edu / ~ scs / Courses / Stat376 / Papers /ConvergeDiagnostics / GelmanRubinStatSci1992 . pdf

Gelman, Andrew & Donald B. Rubin (1995). Avoiding model selection in Bayesian social research. *Sociological Methodology*, S. 165–173. http :// www . stat . columbia . edu / ~ gelman / research/published/avoiding.pdf

Gelman, Andrew & Cosma Rohilla Shalizi (19. Dez. 2010). *Philosophy and the practice of Baye sian statistics in the social sciences*. Oxford Handbook of the Philosophy of the Social Sciences. Hrsg. von Harold Kincaid. Oxford: Oxford University Press. url: http : / / www . stat .columbia.edu/ ~ gelman/research/published/philosophy-_chapter.pdf.

Gelman, Andrew & Cosma Rohilla Shalizi (Feb. 2012). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology 66*(1), S. 8–38. issn: 0007-1102. doi: 10.1111/j.2044-8317.2011.02037.x. url: http://dx.doi.org/10.1111/j.2044-8317.2011.02037.x.

Gelman, Andrew & Cosma Rohilla Shalizi (2013). Philosophy and the practice of Bayesian stati- stics. *British Journal of Mathematical and Statistical Psychology 66*, S. 8–38. url: http: //www.stat.columbia.edu/ ~ gelman-/research/published/philosophy.pdf.

Gelman, Andrew, Daniel Simpson & Michael Betancourt (2017). The Prior Can Often Only Be Understood in the Context of the Likelihood. entropy 19(555). url: http://www.stat. columbia.edu/ ~ gelman/research/published/entropy--19-00555-v2.pdf.

Gelman, Andrew & Hal Stern (2006). The Difference Between »Significant« and »Not Signi- ficant« is not Itself Statistically Significant. *The American Statistician 60*(4), 328–331. url: http://dx.doi.org/10.1198/000313006X152649.

Gelman, Andrew & Francis Tuerlinckx (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics 15*. http://www. stat.columbia.edu/ ~ gelman/research/published-/francis8.pdf, S. 373–390.

Gelman, Andrew & David Weakliem (2009). Of Beauty, Sex and Power. Too little attention has been paid to the statistical challenges in estimating small effects. *American Scientist 97*, 310–316. http://www.stat.columbia.edu/ ~ gelman-/research/published/retropower/final.pdf

Gento, Samuel, Hrsg. (2001a). *La institución educativa. I. Identificadores de calidad.* [Educational institution. I. Identifiers of quality]. Buenos Aires: Docencia.

Gento, Samuel, Hrsg. (2001b). La institución educativa. II Predictores de calidad. [Educational in- stitution. II. Predictors of quality]. Buenos Aires: Docencia.

Gento, Samuel, Hrsg. (2002). *Instuciones educativas para la Calidad Total*. [Educational institutions of total quality]. Madrid: La Muralla.

Gento, Samuel (2014). *Liderazgo en instituciones educativas*. Presentación en la Universidad Veracruzana, Campus Xalapa, Mexico. https ://scielo.conicyt.cl/pdf/perseduc/v57n1/0718-9729-perseduc-57-01-00050.pdf. Mexico.

Gento, Samuel, Günter L. Huber u. a. (2015a). Promoting the Quality of Educational Institutions by Enhancing Educational Leadership. *US-China Education Review B, 5*(4), 215–232. http://www.davidpublisher.org/Public--/uploads/Contribute/5539b2a1a1563.pdf. doi: 10.17265/2161-6248/2015.04.001.

Gento, Samuel, Günter L. Huber et al. (2015b). Promoting the quality of educational institutions by enhancing educational leadership. *US-China Education Review 5*(4), 215–232.
Gerber, Tony [director] (2012). *The Final Word with James Cameron | Titanic: I00*. Documentary movie about the Titanic.
Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernado et al. (Eds.), *Bayesian Statistics* (4th ed.), 169–193. Oxford, UK: Clarendon Press.
Gigerenzer, Gerd (1981). *Messung und Modellbildung in der Psychologie*. München: Reinhardt.
Gigerenzer, Gerd (1991). How to make cognitive illusions disappear: Beyond Heuristics and Biases. *European Review of Social Psychology, 2*, 169–193.
Gigerenzer, Gerd (1993). The Superego, The Ego, and the Id in Statistical Reasoning. In Gideon Keren & Charles Lewis (Eds.), *A Hand book for Data Analysis in the Behavioral Sciences: Methdological Issues* (pp. 311–339). Hillsdale/ New Jersey: Lawrence Erlbaum Assoc.
Gigerenzer, Gerd (2004a). Das Einmaleins der Skepsis. Über den richtigen Umgang mit Zahlen und Risiken. New York: Berliner Taschenbuch Verlag (BvT).
Gigerenzer, Gerd (Nov. 2004b). Mindless statistics. *The Journal of Socio-Economics, 33*(5), 587–606. ISSN: 1053-5357. doi: 10 . 1016 / j . socec . 2004 . 09 . 033. url: http : / / dx.doi.org/10.1016/j.socec.2004.09.033.
Gigerenzer, Gerd (2018). Statistical Rituals: The Replication Delusion and How We Got There. *Advances in Methods and Practices in Psychological Science 1*(2), S. 198–218. https://www.mpib-berlin.mpg.de/pubdata/gigerenzer/ Gigerenzer 2018Statisticalrituals.pdf
Gigerenzer, Gerd & Wolfgang Gaissmaier (2011). Heuristic Decision Making. *Annual Rev. Psychology 62*, 451–482. url: https://library.mpib-berlin.mpg.de/ft/gg/GG_ Heuristic_2011.pdf.
Gigerenzer, Gerd & Ulrich Hoffrage (1995). How to improve Bayesian reasoning without in struction: Frequency formats. *Psychological Review 102*(4), S. 684–704. https://library.mpib- berlin.mpg.de/ft/gg/GG_How_1995.pdf
Gigerenzer, Gerd, Ulrich Hoffrage & A. Ebert (1998). AIDS counselling for low-risk clients. *AIDS CARE 10*(2), 197–211. https://library.mpib-berlin.mpg.de/ft/gg/GG_AIDS_ 1998.pdf
Gigerenzer, Gerd, Stefan Krauss & Oliver Vitouch (2004). The Null Ritual what you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences*,391–408. Thousand Oaks: Sage.
Gigerenzer, Gerd & Julian N. Marewski (2015). Surrogate Science. *Journal of Management 41*(2), 421–440. ISSN: 1557-1211. doi: 10.1177/0149206314547522. url: http://www.dcscience.net/Gigerenzer-Journal-of--Management-2015.pdf.
Gilks, W.R., S. Richardson & D.J. Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall/ CRC.
Gill, Jeff (2007). *Bayesian methods: A social and behavioral sciences approach* (Second Edition) (Chapman and Hall/CRC Statistics in the Social and Behavioral Sciences). Chapman & Hall/CRC. isbn: 1584885629.
Gill, Jeff (2008). *Bayesian methods. A social and behavioral sciences approach*. London: Chapman & Hall/ CRC.
Gilligan, C. Et al. (2003). On the Listening Guide: A Voice-Centred Relational Method. In P.M Camic, J.E. Rhodes, & L. Yardley (Eds.), *Qualitative Research in Psychology: Expanding Perspectives in Methodology and Design*.
Gilligan, C. (1982). *In a different voice. Psychological theory and women's development*. Cambridge, MA, USA: Harvard University Press.
Gilligan, Carol (1982). *The listening guide method of psychological inquiry*. Cambridge, MA: Harvard University Press.
Glaser, Barney G. (1978). *Theoretical sensitivity*. Mill Valey, CA: Sociology Press.
Glaser, Barney G. & Anselm L. Strauss (1967). *Discovery of grounded theory. Strategies for qualitative research*. Aldine Publishing Company.
Glaser, Barney G. & Anselm L. Strauss (1979). Die Entdeckung gegenstandsbezogener Theorie. Eine Grundstrategie qualitativer Sozialforschung. In Christel Hopf & E. Weingarten (Eds.), *Qualitative Sozialforschung*. S. 91–111. Stuttgart: Klett-Cotta.
Glaser, Barney G. & Anselm L. Strauss (1998). *Grounded Theory. Strategien qualitativer Forschung* (Original 1976: The discovery of Grounded Theory). Bern: Hans Huber.
Gödel, Kurt (1931). Über formal unentscheidbare Sätze der »Principia Mathematica« und verwandter Systeme I. *Monatshefte für Mathematik und Physik, 38*, 173–198. http://www.w- k- essler.de/pdfs/goedel.pdf
Goenka, Satya Narayan (1991). *Abendvorträge eines I0-Tage-Kurses in Vipassana-Meditation*. http://www.dhamma.org, http://www.vridhamma.org.Igatpuri, India.
Goenka, Satya Narayan (1997). *The discourse summaries*. Igatpuri, India: V.R.I.
Gönen, Mithat et al. (2005). The Bayesian Two-Sample t Test. *The American Statistician 59*(3)252–257.
Gönen, Mithat et al. (2019). Comparing objective and subjective Bayes Factors for the two- sample comparison: the classification theorem in action. *The American Statistician, 73*(1), 22–31.
Good, I.J. (1979). Studies in the History of Probability and Statistics. XXXVII A. M. Turing's statistical work in World War II. *Biometrika 66*(2), 393–396. doi: 10.1093/biomet/ 66.2.393.MR82c:01049.
Goodman, L.A. & W.H. Kruskal (1954). Measures of association for cross-classification. *Journal of the American Statistical Association 49*, 732–764.
Goodman, Steven N. (1999). Toward evidence-based medical Sstatistics. 1: The P Value Fallacy. *Annals of Internal Medicine 130*(12), 995–1004.

Goodman, Steven N. (2016). Aligning statistical and scientific reasoning. *Science 352*(6290), 1180–1181.
Goodrich, Ben & Jonathan Kropko (2014-06-16). *An Example of mi Usage* [based on earlier versions written by Yu-Sung Su, Masanao Yajima, Maria Grazia Pittau, Jennifer Hill, & Andrew Gelman]. Techn. Ber. https://cran.r- project.org /web/packages/mi/vignettes/mivignette.pdf. CRAN.
Grambsch, P.M. (1994). Simple robust tests for scale differences in paired data. *Biometrika, 81*, 359–372.
Grande, Tilman u. a. (2006). Differential effects of two forms of psychoanalytic therapy: Results of the Heidelberg-Berlin study. *Psychotherapy Research 16*(4), 470–485.
Green, Peter, Catriona MacLeod & Phillip Alday (2019-01-29). *simr: Power Analysis for Generalised Linear Mixed Models by Simulation* [version I.0.5] [Vignettes 'Test examples', 'Power analysis from scratch']. Techn. Ber. CRAN. url: https:// cran.r-project.org/web /packages/simr (visited on 26. 06. 2019).
Green, Peter & Catriona J. MacLeod (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution 7*, 493–498. url: https://besjournals.onlinelibrary.-wiley.com/doi/full/10.1111/2041-210X.12504 (visited on 22. 06. 2019).
Green, Peter [pitakakariki] (2018-01-24). *Couldn't automatically determine a default fixed effect for this model* #!)6 [rsiugzda opened this issue on 24 Jan 20I8]. url: https://github.com/pitakakariki/simr/issues/96 (visited on 26. 06. 2019).
Greene, Brian (2008). *Der Stoff, aus dem der Kosmos ist. Raum, Zeit und die Beschaffenheit der Wirklichkeit*. München: Goldmann.
Greenland, Sander (1998). Induction versus Popper: substance versus semantics. *International Journal of Epidemiology 27*, 543-548. https://www.ph.ucla.edu/epi/faculty/ greenland/Epi204/Greenland1998_InductionvsPopper_IJE.pdf.
Greenland, Sander et al. (2016). Statistical Tests, p values, Confidence Intervals, and Power: A Guide to Misinterpretations. *The American Statistician/ Online Supplement 49*,S. 732–764. https://amstat.tandfonline.com-/doi/suppl /10.1080 /00031305.2016.1154108
Greenwald, Anthony G. u. a. (1996). Effect sizes and p values: What should be reported and what should be replicated. *Psychophysiology 33*, 175–183. https://faculty.washington.edu/agg/pdf/Gwald_Gonz_Har_Guth_P-sychophys_-1996.OCR.pdf
Gregory, Phil C. (2006). *Bayesian Logical Data Analysis for the Physical Sciences. A Comparative Approach with Mathematica® Support*. Cambridge: Cambridge University Press.
Greuel, Fabian (o.J.). *Bayes Klassifikatoren. Methoden des Data Mining*. Natural Lan guage Systems Division, Universität Hamburg. url: https://nats-www.informatik.uni-hamburg.de/pub/Datamining/VeranstaltungsThemen/ BayesKlassifi-katoren.pdf (visited on 22. 05. 2019).
Groeben, Norbert (1986). *Handeln, Tun, Verhalten als Einheiten einer verstehend-erklärenden Psychologie*. Tübingen: Francke Verlag.
Groeben, Norbert & Brigitte Scheele (1977). *Argumente für eine Psychologie des reflexiven Subjekts*. Darmstadt: Steinkopff.
Groeben, Norbert, Diethelm Wahl et al., Hrsg. (1988). *Das Forschungsprogramm Subjektive Theorien. Eine Einführung in die Psychologie des reflexiven Subjekts*. Francke: Tübingen.
Groeben, Norbert & Hans Westmeyer (1981). *Kriterien psychologischer Forschung. Grundfragen der Psychologie*. München: Juventa.
Grund, Simon, Oliver Lüdtke & Alexander Robitzsch (2016). Multiple imputation of multilevel missing data: An introduction to the R package pan. *SAGE Open 6*(4), 1–17. https://www.pedocs.de/volltexte/2017/12692-/pdf/Grund_Luedtke_Robitzsch_2016_Multiple_Imputation_of_multilevel_missing_data.pdf
Gunther, Neil J. (15. Sep. 2013). Laplace the Bayesianista and the Mass of Saturn. url: https://perfdynamics.blogspot.com/-2013/09/laplace-bayesianista-and-mass- of-saturn.html (visited on 22. 05. 2019).
Gürtler, Leo (2005). *Die Rekonstruktion von Innensicht und Aussensicht humorvollen Handelns in Schule und Erwachsenenbildung. Die Bewältigung der Katastrophe — Vipassana¯- Meditation und Humor*. Dissertation an der Sozial- und Verhaltenswissenschaftlichen Fakultät Tübingen. Tübingen: Abteilung Päd. Psych. / Institut für Erziehungswissenschaft / Universität Tübingen.
Gürtler, Leo & Günter L. Huber (2006). The ambiguous use of language in the paradigms in QUAN and QUAL. In Leo Gürtler & Günter L. Huber (Eds.), *Qualitative Research in Psychology (Special Issue on Mixed Methods), 3*(4), 313–328.
Gürtler, Leo & Günter L. Huber (2007). Should we generalize? Anyway, we do it all the time in everyday life. In Leo Gürtler, Mechthild Kiegelmann & Günter L. Huber (Eds.), *Qualitative Psychology Nexus, Vol. 5: Generalization in Qualitative Psychology*, S. 17-35, Tübingen: Verlag Ingeborg Huber.
Gürtler, Leo & Günter L. Huber (2013). *AQUAD 7. Manual — R Integration* [v2, 20I4-0J-22. Softwaremanual. Karlsruhe/ Tübingen.
Gürtler, Leo & Günter L. Huber (2015). Combining qualitative and quantitative analyses. In: Günter L. Huber (Ed.), *Qualitative Psychology Nexus, Vol. 13: New perspectives on qualitative research*, S. 89–106. Tübingen: Center for Qualitative Psychology.
Gürtler, Leo & Günter L. Huber (2016). Computerunterstützte Sequenzanalyse. *Sozialer Sinn 16*(2), 49–70.
Gürtler, Leo & Hartmut-A. Oldenbürger (2005). *Humor aus Schüler/innensicht*. Papier und Vortrag bei dem Symposium des »Forschungsprogramms Subjektive Theorien« 2.–4. März 2005 in Ludwigsburg. http://www.fst-symposium.de/login/guertler_oldenbuerger.pdf. Ludwigsburg.

Gürtler, Leo, Urban M. Studer & Gerhard Scholz (2012). *Tiefensystemik - Band I. Lebenspraxis und Theorie: Wege aus Süchtigkeit finden*. Münster: Monsenstein und Vannerdat Wissenschaft.

Haesebrouck, Tim (2016). The added value of multi-value qualitative comparative analysis. *FQS (Forum Qualitative Research) 17*(1, Art.12). http://nbn-resolving.de/urn:nbn:de:0114-fqs1601129.

Haggbloom, Steven J. u. a. (2002). The 100 most eminent psychologists of the 20th century. *Review of General Psychology) 6*(2). Ranking: https:// www . apa . org / monitor / julaug02 / eminent, https :// www . apa . org / monitor / julaug02 / studyranks, S. 139–152. doi: doi : 10 . 1037 / 1089 - 2680.6.2.139.

Haldane, J.B.S. (1932). A note on inverse probability. *Mathematical Proceedings of the Cambridge Philosophical Society 28*(1), 55–61.

Haller, Heiko & Stefan Krauss (2002). Misinterpretations of significance: a problem students share with their teachers? *Methods of Psychological Research 7*(1), 1–20. https : / / www . metheval . uni - jena . de / lehre / 0405 - ws / evaluationuebung / haller . pdf

Hammerton, M. (1968). Bayesian Statistics and Popper's Epistemology. *Mind. New Series 77*(305), 109–112.

Hampel, F.R. et al., (Eds.) (1986). *Robust Statistics: The Approach based on Influence Functions*. New York: Wiley.

Handl, Andreas (2002). *Multivariate Analysemethoden. Theorie und Praxis multivariater Verfahren unter besonderer Berücksichtigung von S-PLUS*. Berlin: Springer.

Hans-Peter Beck-Bornholdt, & Hans-Hermann Dubben (11. Juni 2005). *Mit an Wahrscheinlichkeit grenzender Sicherheit*. Hamburg: Rowohlt Taschenbuch. ISBN: 3499619024.

Harrell, Frank E. (27. Dez. 2002). *Data for Titanic passengers* [Data obtained from http://biostat.mc.vanderbilt.edu-/DataSets]. Explanations: http : / / biostat . mc . vanderbilt . edu / wiki / pub / Main / DataSets / titanic . html. url: http : //biostat.mc.vanderbilt.edu/wiki/Main/DataSets (visited on 20. 05. 2019).

Hartig, Florian (17. Sep. 2010). *A simple Metropolis-Hastings MCMC in R*. url: https://theoreticalecology.wordpress . com / 2010 / 09 / 17 / metropolis - hastings - mcmc-in-r (visited on 05. 06. 2019).

Hastings, W.K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika 57*, 97–109.

Hathaway, S.R. & J.C. McKinley (1940). A multiphasic personality schedule (Minnesota): 1. Costruction of the schedule. *Journal of Psychology 10*, 249–254.

Hattie, John (2009). *Visible learning*. London: Routledge.

Hausser, Jean & Korbinian Strimmer (2009). Entropy Inference and the James-Stein Estimator, with Application to Nonlinear Gene Association Networks. *Journal of Machine Learning Research 10,* 1469–1484. R-package: http://www.strimmerlab.org/software/entropy. url: http://www.jmlr.org/papers/volume10/hausser09a-/hausser09a.pdf (visited on 05. 06. 2019).

Havighurst, Robert J. (1953). *Human development and education*. New York: Longmans.

Hays, William L. (1974). *Statistics for the Social Sciences* [2nd ed. with corrections, Ist ed. 1963]. London: Holt, Rinehart & Winston.

Head, Megan L. et al. (2015). The extent and consequences of p-hacking in science. *PLoS Biology 13*(3). https://journals.plos.org/plosbiology/article?id = 10.1371/ journal.pbio.1002106.

Heatherton, Todd F. & Joel Lee Weinberger (Eds.) (1994). *Can personality change?* Washington, DC: American Psychological Association.

Heckhausen, Heinz (1989). *Motivation und Handeln*. Berlin: Springer.

Heidelberger, P. & P.D. Welch (1981). A spectral method for confidence interval generation and run length control in simulations. *Comm. ACM 24*, 233–245.

Heidelberger, P. & P.D. Welch (1983). Simulation run length control in the presence of an initial transient. *Opns Research 31*, 1109–1144.

Heisenberg, Werner (1927). Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. *Zeitschrift für Physik 43*(3), 172–198. https : / / web . archive . org / web / 20130510070844, http : / / osulibrary . oregonstate . edu / specialcollections / coll / pauling / bond / papers / corr155 . 1 . html.

Held, Josef (1994). *Praxisorientierte Jugendforschung: Theoretische Grundlagen, methodische Ansätze, exemplarische Projekte*. Hamburg: Argument-Verlag.

Held, Josef, H.-W. Horn & A. Marvakis (1994). *Gespaltene Jugend. Politische Orientierungen jugendlicher ArbeitnehmerInnen und ihre subjektiven Begründungen im Kontext gesellschaftlicher Veränderungen*. Opladen: Leske + Budrich.

Held, Josef & A. Marvakis (1992). Empirische Jugendforschung und ihr Verhältnis zur politischen Bildung. Lernwidersprüche und pädagogisches Handeln. In K.-H. Braun & K. Wetzel (Eds.), *Bericht von der 6. internationalen Ferienuniversität Kritische Psychologie* (S. 243–256). Marburg: Verlag Arbeit & Gesellschaft.

Held, Leonhard & Manuela Ott (2016). How the Maximal Evidence of P-Values Against Point Null Hypotheses Depends on Sample Size. *The American Statistician 70*(4), 335–341. doi: 10.1080/00031305.2016.1209128.

Held, Leonhard & Manuela Ott (2018). On p-values and Bayes Factors. *Annual Review of Statistics and Its Application 5*(1), 393–419.

Hempel, C.G. & P. Oppenheim (1948). Studies in the logic of explanation. *Philosophy of Science, 15*, 135–175.

Hennig, Christian & Mahmut Kutlukaya (2007). Some thoughts about the design of loss functions. *Revstat — Statistical Journal 5*(1), 19–39. https : / / www . ine . pt / revstat / pdf / rs070102.pdf.

Herrmann, Andrea Monika & Lasse Cronqvist (2009). When dichotomisation becomes a pro- blem for the analysis of middle-sized datasets. *International Journal of Social Research Methodology 12*(1), 33–50.. https :// www .andrea-herrmann.eu/publications/Herrmann_Cronqvist IJSRM 2009 When_Dichotomisation_becomes_a_-Problem.pdf

Hicks, Tyler, Liliana Rodriguez-Campos & Jeong Hoon Choi (2017). Bayesian Posterior Odds Ratios: Statistical Tools for Collaborative Evaluations. *American Journal of Evaluation 39*(2), 278–289.

Hildenbrand, Bruno (1996). *Methodik der Einzelfallstudie: Theoretische Grundlagen, Erhebungs- und Auswertungsverfahren, vorgeführt an Fallbeispielen*. Studienbrief in drei Bänden. Hagen: Fernuniversität Hagen.

Hildenbrand, Bruno (1999). *Fallrekonstruktive Familienforschung. Anleitungen für die Praxis*. 2. Auflage 2005. Wiesbaden: Verlag für Sozialwissenschaften.

Hildenbrand, Bruno (2005). *Einführung in die Genogrammarbeit*. Heidelberg: Carl-Auer Systeme.

Hildenbrand, Bruno (8. März 2006a). *Fallrekonstruktive Familienforschung: Theorie und Grundlagen*. Workshop zur Einzelfallforschung auf der 3. Tagung »Systemische Forschung in Therapie, Pädagogik und Organisationsentwicklung«. Heidelberg.

Hildenbrand, Bruno (2006b). Resilienz in sozialwissenschaftlicher Perspektive. In: Rosmarie Welter-Enderlin & Bruno Hildenbrand (Eds.), *Resilienz — Gedeihen trotz widriger Umstände*, S. 20–27. Heidelberg: Carl-Auer Systeme,

Hildenbrand, Bruno (2018). *Genogrammarbeit für Fortgeschrittene: Vom Vorgegebenen zum Aufgegebenen*. Heidelberg: Carl-Auer Systeme.

Hino, Airo (2009). Time-Series QCA Studying Temporal Change through Boolean Analysis. *Sociological Theory and Methods 24*(2), 247–265. url: https://www.jstage.jst.go. jp/article/ojjams/24/2/24_2_247/_pdf (visited on 20. 03. 2019).

Hoenig, John M. & Dennis M. Heisey (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician 55*(1), 19–24.

Hoff, Peter D. (2009). *A first course in Bayesian Statistical Methods*. Dordrecht: Springer.

Hoffman, Matthew D. & Gelman Andrew (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research 15*, 1593–1623. url: http://jmlr.org/papers-/v15/hoffman14a.html (visited on 05. 06. 2019).

Honaker, James, Gary King & Mathew Blackwell (2011). Amelia II: A Program for Missing Data. J*ournal of Statistical Software 45*(7), 1–47. https : / / gking . harvard . edu / files / gking/files/amelia_jss.pdf.

Honaker, James, Gary King & Mathew Blackwell (2018-05-07). *AMELIA II: A Programfor Missing Data* [version I.J.5] [Vignette]. Techn. Ber. CRAN. url: https://cran.r-project. org/web/packages/Amelia/vignettes/amelia.pdf (visited on 28. 06. 2019).

Honorton, C. & D.C. Ferrari (1989). Future telling: A meta-analysis of force-choice precognition experiments, 1935–1987. *Journal of Parapsychology, 53*, 281–308.

Hossenfelder, Sabine (3. Feb. 2012). *The Free Will Function*. Unpublished manuscript [physics.hist-ph]. url: https://arxiv.org/abs/1202.0720 (visited on 23. 05. 2019).

Hossenfelder, Sabine (10. Jan. 2016). *Free will is dead, let's bury it*. url: https : / / backreaction . blogspot . com / 2016 / 01 / free - will - is - dead - lets - bury - it.html (visited on 23. 05. 2019).

Hossenfelder, Sabine (2018). *Das hässliche Universum. Warum unsere Suche nach Schönheit die Phyik in die Sackgasse führt*. Übersetzt von Gabriele Gockel & Sonja Schuhmacher. Frankfurt am Main: Fischer.

Hossenfelder, Sabine (2. Mai 2019). *How to live without free will*. url: http://backreaction. blogspot.com/2019/05-/how-to-live-without-free-will.html (visited on 23. 05. 2019).

Hoyda, Joseph J., Alyssa Counsell & Robert A. Cribbie (2019). Traditional and Bayesian Approaches for Testing Mean Equivalence and a Lack of Association. *The Quantitative Methods for Psychology 15*(1), 12–24. url: https://www.tqmp.org-/egularArticles/vol15-1/p012/p012.pdf (visited on 29. 06. 2019).

Hubbard, Raymond (Juni 2004). Alphabet Soup. *Theory and Psychology 14*(3), 295–327. ISSN: 1461-7447. url: http://dx.doi.org/10.1177/0959354304043638.

Hubbard, Raymond & M. J. Bayarri (Aug. 2003). Confusion Over Measures of Evidence (p's) Versus Errors (?'s) in Classical Statistical Testing. *The American Statistician 57*(3), 171–178. issn: 1537-2731. doi: 10.1198/0003130031856. url: http://dx.doi.org/10. 1198/0003130031856.

Hubbard, Raymond & R. Murray Lindsay (Feb. 2008). Why P Values Are Not a Useful Measure of Evidence in Statistical Significance Testing. *Theory and Psychology 18*(1), 69–88. ISSN: 1461-7447. url: http://dx.doi.org-/10.1177/0959354307086923.

Hubbard, Raymond & Patricia A. Ryan (Okt. 2000). Statistical Significance with Comments by Editors of Marketing Journals. Educational and Psychological Measurement 60(5), 661–681. ISSN: 1552-3888. doi: 10.1177/0013164400605001. url: http://dx.doi.org/10.1177/0013164400605001.

Huber, Anne A. (Ed.). (2007). *Wechselseitiges Lehren und Lernen (WELL) als spezielle Formen Kooperativen Lernens*. Berlin: Logos Verlag.

Huber, Günter L (1992). Qualitative Analyse mit Computerunterstützung. In Günter L. Huber (Ed.), *Qualitative Analyse. Computereinsatz in der Sozialforschung* (S. 115–175). München: Oldenbourg.

Huber, Günter L. & Leo Gürtler (2012). *AQUAD Sieben. Manual zur Software AQUAD 7*. (1. Auflage 2003). http :// www . aquad . de / materials / manual _ aquad7 / manual - d. pdf. Tübingen: Softwarevertrieb Günter Huber (Original-veröffentlichung: Tübingen: Ingeborg Huber Verlag).

Huber, Günter L., Leo Gürtler & Samuel Gento (2018). La aportación de la estadística exploratoria al análisis de datos cualitativos. *Perspectiva Educacional. Formación de Profesores 57*(1), 50–69.

Huber, Günter L. & S. Kenntner (1986). *Struktur biographischer Selbst-Schemata*. Techn. Bericht, Deutsche Forschungs-gemeinschaft, Projekt Hu 348/2-3. Tübingen: Universität Tübingen.

Huber, Günter L. (unter Mitarbeit von Leo Gürtler) (2019). *AQUAD — Software zur Analyse qualitativer Daten (Version 7)*. http://www.aquad.de. Softwarevertrieb Günter Huber. Tübingen.

Huber, Peter J. (Ed.) (1981). *Robust Statistics*. New York: Wiley.

Hubert, Lawrence (1974). Approximate evaluation techniques for the single-link and complete- link hierarchical clustering procedure. *Journal of the American Statistical Association 69*, 698–704.

Hughes, Jeffrey (2015-01-29). *Evaluating the R-Index and the P-Curve*. http : / / disjointedthinking . jeffhughes . ca / 2015 / 01 / evaluating - r - index - p - curve/.

Humphreys, Macartan & Alan M. Jacobs (2015). Mixing Methods: A Bayesian Approach. *American Political Science Review 109*(4), 653–673. http : / / www . columbia . edu / ~mh2245 / papers1/BIQQ.pdf.

Hussy Walter & Möller, H. (1994). Hypothesen. In Theo Hermann & W.H. Tack (Eds.), *Enzyklopädie der Psychologie. Themenbereich B: Methdologie und Methoden. Serie I: Forschungsmethoden der Psychologie. Band I: Methodologische Grundlagen der Psychologie*, S. 653–673. Göttingen: Hogrefe.

Hussy, Walter (1986). *Denkpsychologie. Ein Lehrbuch. Band 2: Schlußfolgern, Urteilen, Kreativität, Sprache, Entwicklung, Aufmerksamkeit*. Stuttgart: Kohlhammer.

Inzunsa Cazares, S. (2018). Razonamiento estadístico de estudiantes universitarios sobre el análisis de datos en un ambiente computacional. *Bolema: Boletim de Educação Matemática 28*(50), 1262–1286. http://dx.doi.org-/10.1590/1980-4415v28n50a13,.

Ionides, Edward L. et al. (2017). Letters to the Editor: Response to the ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician 71*(1), 88–89.

Iverson, Geoffrey J. et al. (2009). Prep: An agony in five Fits. *Journal of Mathematical Psychology 53*, 195–202.

Jackson, Ronny L. (16. Jan. 2018). *Physician to the president. The White House. Memorandum*. url: https://de.scribd.com/document/369309002/Summary-of-Physical-Exam-for-President-Trump-12-Jan-2018 (visited on 16. 01. 2018).

Jakobson, Thorsten et al. (2007). Ergebnisse analytischer Langzeitpsychotherapien bei spezifischen psychischen Störungen: Verbesserungen in der Symptomatik und in interpersonellen Beziehungen. *Zeitschrift fur Psychosomatische Medizin und Psychotherapie 53*(2), 87–110. https://www.researchgate.net-/publication/273662107_Ergebnisse_analytischer_ Langzeitpsychotherapien _ bei _ spezifischen _ psychischen _ Storungen _ Verbesserungen _ in _ der _ Symptomatik _ und _ in _ interpersonellen _ Beziehungen.

Janis, Irving L. (1971). Groupthink. *Psychology Today Magazine 5*(6), 84–90. https : / / web . archive . org / web / 20100401033524, pdf: http : / / apps . olin . wustl . edu / faculty/macdonald/GroupThink.pdf.

Jaynes, Edwin Thompson (1958). *How does the brain do plausible reasoning* [first appeared as laboratory report, short version published in Erickson, G. J. and Smith, C. R. (Eds.). Maximum-Entropy and Bayesian Methods in Science and Engineering, I, Kluwer, Dordrecht, p.I. extended laboratory report. Long original version: https://bayes.wustl.edu-/etj/articles/how.does.the.brain.orig.pdf, shorter published version: https://-bayes.wustl.edu/etj/articles/brain.pdf. Stanford, California: Microwave Laboratory & Department of Physics, Stanford University.

Jaynes, Edwin Thompson (1963). *Information Theory and Statistical Mechanics* (Notes by the Lecturer) [Statistical Physics 3]. Lectures from Brandeis Summer Institute 1962. https : / / bayes.wustl.edu/etj/articles/brandeis.pdf. New York: WA: Benjamin, Inc.

Jaynes, Edwin Thompson (1968). *Prior Probabilities*. IEEE Trans. on Systems Science and Cybernetics sec-4(3), 227–241. url: https://bayes.wustl.edu/etj/articles/prior. pdf (visited on 24. 05. 2019).

Jaynes, Edwin Thompson (1976). Confidence Intervals vs. Bayesian Intervals. In: W.L. Harper & C.A. Hooker (Eds.) *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science* S. 175–257. Dor drecht/ NL: D. Reidel,. url: http : / / bayes . wustl.edu/etj/articles/confidence.pdf (visited on 05. 06. 2019).

Jaynes, Edwin Thompson (1978). Where do we stand on Maximum Entropy? [original: presentation at the Maximum Entropy Formalism Conference, Massachusetts Institute of Technology, May 2–4, 1978.] In: Raphael D. Levine & Myron Tribus (Eds.), *The Maximum Entropy Formalism*, S. 15-118.. Cambridge/MA: MIT Press. url: https://bayes.wustl.edu/etj/articles/stand.on.entropy.pdf (visited on 05. 06. 2019).

Jaynes, Edwin Thompson (1983). *Paperson Probability, Statistics and Statistical Physics*. Dordrecht/ NL: D. Reidel Publishing Co.

Jaynes, Edwin Thompson (1986a). Bayesian Methods: General background. An Introductory Tutorial. In James H. Justice (Ed.), *Maximum-Entropy and Bayesian Methods in Applied Statistics*, S. 1–25. Cambridge: Cambridge University Press.

Jaynes, Edwin Thompson (1986b). Monkeys, Kangaroos, and N [Presented at the Fourth Annual Workshop on Bayesian/Maximum Entropy Methods, University of Calgary, August 1984. Published in the Proceedings. The present version was revised and corrected, to clarify some arguments and include new understanding, in June 1994. In James

H. Justice (Ed.), *Maximum Entropy and Bayesian Methods in Applied Statistics*. https://bayes.wustl.edu/etj-/articles/cmonkeys.pdf. Cambridge: Cambridge University Press.

Jaynes, Edwin Thompson (1988a). Probablity Theory as Logic. Talk presented at the Ninth Annual Workshop on Maximum Entropy and Bayesian Methods, Dartmouth College, New Hampshire, August 1988. In Paul F. Fougere, (Ed.), *Maximum Entropy and Bayesian Methods*. Dordrecht: Kluwer Academic Publishers. The present version was substantially revised, corrected, and extended. Techn. Ber. https:// bayes.wustl.edu/etj/articles/prob.as.logic.pdf.

Jaynes, Edwin Thompson (1988b). The Relation of Bayesian and Maximum Entropy. In G.J. Erickson & C.R. Smith (Eds.), *Maximum-Entropy and Bayesian Methods in Science and Engineering* (Vol. I), S. 25–29. Dordrecht: Kluwer Academic Publishers

Jaynes, Edwin Thompson (2003). *Probability theory: The logic of science*. Edited by G. Larry Bretthorst. Cambridge: Cambrige University Press.

Jaynes, Edwin Thompson (1957a). Information Theory and Statistical Mechanics [Reprint]. *The Physical Review 106*(4), 620–630. url: https : / / bayes . wustl . edu / etj / articles/theory.1.pdf (visited on 22. 05. 2019).

Jaynes, Edwin Thompson (1957b). Information Theory and Statistical Mechanics. II [Reprint]. *The Physical Review 108*(2), 171–190. url: https : / / bayes . wustl . edu / etj / articles/theory.2.pdf (visited on 22. 05. 2019).

Jaynes, Edwin Thompson (1996-08-07). Probability in quantum theory. Techn. Ber. A revised and extended version of a paper presented at the Workshop on Complexity, Entropy, and the Physics of Information, Santa Fe, New Mexico, May 29–June 2, 1989. The original version is in W. H. Zurek (Ed.), *Complexity, Entropy and the Physics of Information*. Reading, MA: Addison Wesley Publishing Co. http://ba s.wustl.edu/etj/articles-/prob.as.logic.pdf.

Jeffreys, Harold (1931). *Scientific Inference*. (3rd ed. 1973). Cambridge: Cambridge University Press.

Jeffreys, Harold (1935). Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Phil. Society. 31*, 203-222. https://www.uvm.edu/pdodds/files/papers/others/1935/jeffreys1935a.pdf.

Jeffreys, Harold (1939/1961). *Theory of Probability*. (3rd ed.; 1st ed. 1939). London: Oxford University Press.

Joachim (n.d.). *Predictive Mean Matching Imputation (Theory and Example in R). Statistical Programming*. url: https :// statistical - programming . com / predictive - mean - matching-imputation-method (visited on 23. 06. 2019).

John, Leslie K., George Loewenstein & Drazen Prelec (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science 23*(5), 524–532. https : / / www . cmu . edu / dietrich / sds / docs / loewenstein / MeasPrevalQuestTruthTelling.pdf.

Johnson-Laird, Philip Nicholas (1993). *Human and machine thinking*. Hillsdale/ NJ: Lawrence Erlbaum Associates.

Jopt, Uwe-Jörg (1978). Warum manche Schüler 'faul' sind: Die attributionstheoretische Vernünftigkeit des schulischen Anstrengungsverzichts. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie 10*(4), 315–327.

Joseph, Maxwell B. (8. Sep. 2013). *Animating the Metropolis algorithm. A homemade Metropolis al gorithm animation using R and the animation package*. url: https://mbjoseph.github.io/posts/2018-12-25-animating-the-metropolis--algorithm.

Jurecková, Jana (1984). 21 M-, L- and R-estimators. *Handbook of Statistics 4*, 463–485.

Kabacoff, Robert I. (2017). *Regression Diagnostics*. url: https://www.statmethods.net/ stats/rdiagnostics.html (visited on 21. 05. 2019).

Kabat-Zinn, Jon (2005). Gesund durch Meditation. Das große Buch der Selbstheilung. Frankfurt a. Main: O.W. Barth.

Kahan, Brennan C., Sunita Rehal & Suzie Cro (2015). Risk of selection bias in randomised trials. *Trials 16*, 405. url: https :// www.ncbi.nlm.nih.gov/pmc/articles/PMC4566301 (visited on 27. 06. 2019).

Kahnemann, Daniel (26. Sep. 2012). *A proposal to deal with questions about priming effects [An open letter to the scientific community]*. Copy hosted on Decision Science News: http : / / www . decisionsciencenews . com / 2012 / 10 / 05 / kahneman - on - the - storm - of - doubts - surrounding - social - priming - research, *full discussion*: https : / / files . osf . io / v1 / resources / cc2xq / providers / osfstorage / 58d93c926c613b0048a068e6?action=-download&direct&version=1. *Nature*. url: https://www.nature.com/polopoly_fs/7.6716.1349271308!/suppinfoFile/Kahneman%5C%20Letter.pdf.

Kaiser, Henry F. (1960). Directional Statistical Decisions. *Psychological Review 67*(3), 160– 167.

Karae Jacobi (5. Dez. 2015). *What does the following quotation from Macbeth mean in modern English: »By the pricking of my thumbs, something wicked this way comes?*. url: https://www. enotes . com / homework - help / what - following - quotation - from - macbeth - mean-modern-580614 (visited on 22. 03. 2019).

Karn, Ujjwal (20. Jan. 2018). *xda: R package for exploratory data analysis*. url: https://github. com/ujjwalkarn/xda.

Kass, Robert E. (1993). Bayes Factors in Practice. *Journal of the Royal Statistical Society. Series D (The Statistician). Special Issue: Conference on Practical Bayesian Statistics,* 42(5), 551–560.

Kass, Robert E. & Adrian E. Raftery (1993). *Bayes Factors and Model Uncertainty* [Technical Report No.254, March 1993]. Techn. Ber. https://pdfs.semanticscholar.org/42d6/ 71ae17a611ac474cb39f59f4cf31f65b51ef.pdf. Seattle/ Washington 98195 USA: Department of Statistics, GN-22, University of Washington.

Kass, Robert E. & Adrian E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*(430), 773–795. https://www.stat.washington.edu/raftery/Research/PDF/kass1995.pdf.

Kass, Robert E. & Larry Wasserman (1996). The Selection of Prior Distributions by Formal Ru les. *Journal of the American Statistical Association 91*(435), 1343–1370. https://www.phil.vt.edu/dmayo/PhilStatistics/Kass%20and%20Wasserma-%201996%20Selection%20of%20prior%20distributions%20by%20formal-%20rules.pdf.

Kelle, Udo & Susann Kluge (1999). *Vom Einzelfall zum Typus. Fallvergleich und Falkontrastie rung in der qualitativen Sozialforschung*. Opladen: Leske + Budrich.

Kelly, George A. (1955). *The psychology of personal constructs*. New York: Norton.

Keng, Brian (27. Jan. 2017). *Maximum Entropy Distributions*. url: http://bjlkeng.github. io/posts/maximum-entropy–distributions (visited on 05. 06. 2019).

Kenntner, Samuel, Fischer, Peter & Huber, Günter L. (1988). *Mehr als eine Folge von Lebensereignissen: Autobiographische Darstellungen als Verarbeitung von Selbstschemata*. Bd. Poster zum 36. Kongress der Deutschen Gesellschaft für Psychologie, Berlin, Oktober 1988.

Kiegelmann, Mechthild et al. (2000). Das Zentrum für Qualitative Psychologie an der Universität Tübingen [24 paragraphs]. *Forum Qualitative Sozialforschung [On-line Journal] 1*(2), Art.14. 2020-07-22, http://nbn-resolving.de-/urn:nbn:de:0114-fqs0002143.

Kievit, Rogier A. et al. (2013). Simpson's paradox in psychological science: a practical guide. *Frontiers in Psychology 4*. http://dx.doi.org/10.3389/fpsyg.2013.00513. doi: 10.3389/fpsyg.2013.00513.

Kimball, Allyn W. (1957). Errors of the Third Kind in Statistical Consulting. *Journal of the American Statistical Association 52*(278), 133–142.

Kish, Leslie (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Kleiber, Christian & Zeileis, Achim (2016). Visualizing Count Data Regressions Using Rootograms. *The American Statistician 70*(3), 296–303. https :// arxiv . org / abs / 1605 . 01311.

Klein, Richard A. et al. (2014). Investigating Variation in Replicability. *Social Psychology 45*(3), 142–152. doi: 10.1027/1864-9335/a000178. eprint: https://doi.org/10.1027/ 1864-9335/a000178. url: https://doi.org/10.1027-/1864-9335/a000178.

Kluge, Susann (Jan. 2000). Empirisch begründete Typenbildung in der qualitativen Sozialforschung. *Forum Qualitative Sozialforschung 1*. http : / / www . qualitative - research.net/fqs-texte/1-00/1-00kluge-d.htm.

Koehrsen, Will (28. Jan. 2018). *Overfitting vs. Underfitting: A Complete Example. Exploring and solving a fundamental data science problem*. url: https://towardsdatascience.com/ overfitting- vs- underfitting- a- complete- example-d05dd7e19765 (visited on 29. 05. 2019).

Korzybski, Alfred (1933). *A Non-Aristotelian System and its Necessity for Rigour in Mathematics and Physics*. Paper presented at the meeting of the American Association for the Advancement of Science in New Orleans, Louisiana on December 28, 1931. Reprinted in Science and Sanity, 1933, pp.747–761.

Korzybski, Alfred (1941). *Science and sanity. An Introduction to Non-Aristotelian Systems and General Semantics* (first published 19333). Lakeville, Conn.: International Non-aristotelian Library Publishing Co.

Kracauer, Siegfried (1952). The challenge of qualitative content analysis. *Public Opinion Quarterly*, *16*, 631-642.

Kraimer, Klaus, Hrsg. (2000). Die Fallrekonstruktion. Sinnverstehen in der sozialwissenschaftlichen Forschung. Frankfurt am Main: Suhrkamp.

Krämer, Walter (1998). *Denkste. Trugschlüsse aus der Welt der Zahlen und des Zufalls*. Piper. ISBN: 9783492224437.

Krämer, Walter (2000). *So lügt man mit Statistik*. Piper. ISBN: 9783492230384.

Krapp, A., Hofer, M. & Prell, V. (1982). *Forschungs-Wörterbuch. Grundbegriffe zur Lektüre wissenschaftlicher Texte*. München: Urban & Schwarzenberg.

Krause, Robert (17. Jan. 2019). *Multiple Imputation for RSiena*. url: http://www.stats.ox.ac.uk/∼snijders/siena/-AdSUMMissingDataMD.html (visited on 30. 07. 2020).

Krause, Robert W., Mark Huisman & Tom A.B. Snijders (2018). Multiple Imputation for Longitudinal Network Data. *Statistica Applicata — Italian Journal of Applied Statistics 30*(1), 33–57. url: http://sa-ijas.stat.unipd.it/sites-/sa-ijas.stat.unipd.it/files/10.26398-IJAS.0030-002.pdf (visited on 23. 06. 2019).

Krauth, Joachim, Hrsg. (1995). *Testkonstrunktion und Testtheorie*. Weinheim: Beltz/ PVU.

Kriegeskorte, Nikolaus, Jerzy Bodurka & Peter Bandettini (2008). Artifactual time-course correlations in echo-planar fMRI with implications for studies of brain function. International Journal of Imaging Sytsems and Technology 18(5–6), S. 345–349.

Krook, Mona Lena (2010). Women's representation in parliament: A qualitative comparative analysis. *Political Studies 58*(5), 886–908. https :// www . researchgate . net / profile / Ivo _ De _ Sousa / post / Doing _ Qualitative _ Comparative _ Analysis _ on _ large _ data _ sets / attachment / 59d6271479197b80779855d9 / AS : 324512598429696 @ 1454381240961/download/QCA + Women%27s + Representation + in + Parliament- + A + Qualitative + Comparative + Analysis.pdf.

Kruschke, John K. (2011a). Bayesian Assessment of Null Values Via Parameter Estimation and Model Comparison. *Perspectives on Psychological Science 6*(3), 299–312. url: http://dx. doi.org/10.1177/1745691611406925.

Kruschke, John K. (2011b). Bayesian Assessment of Null Values Via Parameter Estimation and Model Comparison. *Perspectives on Psychological Science 6*(3), 299–312.

Kruschke, John K. (6. Sep. 2012a). *Posterior predictive check can and should be Bayesian*. url: http :// doingbayesiandataanalysis . blogspot . com / 2012 / 09 / posterior - predictive-check-can-and.html (visited on 03. 06. 2019).

Kruschke, John K. (16. Apr. 2012b). *Why to use highest density intervals instead of equal tailed intervals*. url: http://doingbayesiandataanalysis.blogspot.com/2012/04/why- to-use-highest-density-intervals.html (visited on 29. 05. 2019).

Kruschke, John K. (2013a). Bayesian Estimation Supersedes the t Test. *Journal of Experimental Psychology: General 142*(2), S. 573–603. http : / / www . indiana . edu / %7Ekruschke / articles / Kruschke2013JEPG.pdf.

Kruschke, John K. (8. Aug. 2013b). *How much of a Bayesian posterior distribution falls inside a region of practical equivalence (ROPE)*. url: https : / / doingbayesiandataanalysis . blogspot.com/2013/08/how- much- of- bayesian-posterior.html (visited on 29. 05. 2019).

Kruschke, John K. (2013c). Posterior predictive checks can and should be Bayesian: Comment on Gelman and Shalizi, 'Philosophy and the practice of Bayesian statistics'. *British Journal of Mathematical and Statistical Psychology 66*(1), 45–56. url: http://dx.doi.org/10. 1111/j.2044-8317.2012.02063.x.

Kruschke, John K. (9. Apr. 2015a). *Bayes factors for tests of mean and effect size can be very different*. url: https://doingbayesiandataanalysis.blogspot.com/2015/04/bayes- factors-for-tests-of-mean-and.html (visited on 28. 05. 2019).

Kruschke, John K. (11. Jan. 2015b). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan* (2nd ed.). Oxford: Elsevier LTD, ISBN: 0124058884.

Kruschke, John K. (26. Jan. 2015c). *Institutionalized publication thresholds, p values, and XK-CD*. url: http : / / doingbayesiandataanalysis . blogspot . com / 2015 / 01 / institutionalized-publication.html (visited on 29. 05. 2019).

Kruschke, John K. (27. Dez. 2015d). *Lessons from Bayesian disease diagnosis: Don't over-interpret the Bayes factor, VERSION 2*. url: https://doingbayesiandataanalysis.blogspot. com / 2015 / 04 / bayes - factors - for - tests - of - mean - and . html (visited on 28. 05. 2019).

Kruschke, John K. (21. Dez. 2016a). *Bayesian assessment of null values*. url: http : / / doingbayesiandataanalysis.blogspot.com/2016/12/bayesian-assessment- of-null-values.html (visited on 24. 05. 2019).

Kruschke, John K. (11. Juli 2016b). *MCMC effective sample size for difference of parameters (in Bayesian posterior distribution)*. url: https://doingbayesiandataanalysis.blogspot. com/2016/07/mcmc-effective-sample-size-for.html (visited on 11. 07. 2016).

Kruschke, John K. (22. Okt. 2016c). *Posterior predictive distribution for multiple linear regression*. url: http : / / doingbayesiandataanalysis . blogspot . com / 2016 / 10 / posterior-predictive-distribution-for.html (visited on 03. 06. 2019).

Kruschke, John K. (2016d). *Programs used in the book «Doing Bayesian Data Analysis (2nd ed.)»* [latest version: 20I6-0J-I0]. url: https : / / sites . google . com / site / doingbayesiandataanalysis/software-installation.

Kruschke, John K. (16. Feb. 2017a). *Equivalence testing (two one-sided test) and NHST compared with HDI and ROPE*. url: https://doingbayesiandataanalysis.blogspot.com/ 2017 / 02 / equivalence - testing - two - one - sided - test . html (visited on 24. 05. 2019).

Kruschke, John K. (11. Juni 2017b). *Posterior distribution of predictions in multiple linear regression*. url: https : / / doingbayesiandataanalysis . blogspot . com / 2017 / 06 / posterior-distribution-of-predictions.html (visited on 03. 06. 2019).

Kruschke, John K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practice in Psychological Science 1*(2), 270–280.

Kruse, Rudolf, Jörg Gebhardt & Frank Klawonn (1994). *Foundations of fuzzy systems*. 2. enlarged ed. Chichester: John Wiley & Sons. http :// fuzzy . cs . ovgu . de / studium / fuzzy / txt / fsbook . pdf.

Krynski, Tevye R. & Joshua B. Tenenbaum (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General 136(*3, 430–450. https :// web . mit . edu/ ~ cocosci/Papers/krynski-tenenbaum--- jepgen07.pdf

Kuhn, Thomas S. (1973). *Die Struktur wissenschaftlicher Revolutionen*. Frankfurt am Main: Suhrkamp.

Kullback, Solomon & Richard A. Leibler (1951). On Information and Sufficiency. *Annals of Mathematical Statistics 22*(1), 79–86. https :// projecteuclid . org / euclid . aoms / 1177729694.

Lakens, Daniël (Apr. 2015). Comment: What p-hacking really looks like: A comment on Masicampo and LaLande (2012). *Quarterly Journal of Experimental Psychology 68*(4), 829–832. ISSN: 1747-0226. doi: 10.1080/17470218- .2014.982664. url: http://dx.doi.org/10.1080/17470218.2014.982664.

Lakens, Daniël (2017a). Evquivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science 8*(4), 355–362. url: https:/ / journals . sagepub . com / doi / 10 . 1177 / 1948550617697177 (visited on 29. 06. 2019).

Lakens, Daniël (29. Jan. 2017b). *Examining Non-Significant Results with Bayes Factors and Equivalence Tests*. url: http :// daniellakens . blogspot . com / 2017 / 01 / examining - non-significant-results-with.html (visited on 29. 05. 2019).

Lakens, Daniël (12. Feb. 2017c). *ROPE and Equivalence Testing: Practically Equivalent?* url: http : / / daniellakens . blogspot . com / 2017 / 02 / rope - and - equivalence - testing.html (visited on 24. 05. 2019).

Lakens, Daniël (2018-08-03). TOSTER: *Two One-Sided Tests (TOST) Equivalence Testing* [version 0.3.4] [Vignette 'Introduction to TOSTER']. Techn. Ber. Vignette: https : / / cran . r - project.org/web/packages/TOSTER/vignettes-/IntroductionToTOSTER.html. CRAN. url: https://cran.r- project.org/web/packages/TOSTER (visited on 29. 06. 2019).

Lakens, Daniël, Neil McLatchie et al. (2018). Improving Inferences About Null Effects With Bayes Factors and Equivalence Tests. *The Journals of Gerontology. Series B*. R-Code: https :// osf .io/ hwgnj. url: https :// academic . oup . com / psychsocgerontology /advance - article / doi / 10 . 1093 / geronb / gby065 / 5033832 (visited on 29. 06. 2019).

Lakens, Daniël, Scheel, Anne M. & Isager, Peder M. (2018). Evquivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science 2*(1), 259–269. url: https : / / journals . sagepub . com / doi / 10 . 1177 /2515245918770963 (visited on 29. 06. 2019).

Laktos, Imre (1978). *The methodology of scientific research programmes* (Edited by John Worrall and Gregory Currie). Cambridge: Cambridge University Press.

LaMont, Colin Harry & Paul A. Wiggins (2016-10). *The Lindley paradox: The loss of resolution in Bayesian inference*. https://www.researchgate.net/publication/309572961_The_Lindley_paradox.The_loss_of_resolution_in_Bayesian_inference.

Laplace, Pierre-Simon (1814). *Essai philosophique sur les probabilités* [A Philosophical Essay on Pro babilities]. translated from the 6th French edition by Frederick Wilson Truscott and Frederick Lincoln Emory. New York: John Wiley and Sons, 1902, Dover Publications edition, New York, 1951, has same pagination. New York: Dover Publishing.

Laplace, Pierre-Simon (1843). *Exposition delathéorie des chances et des probabilités* [Paris: engl. Über setzung 1951]. New York: Dover Publishing.

Leeper, Thomas J. (2009). *Introduction to R: Course materials for teaching R* [latest version: 20IJ]. url: https://thomasleeper.com/Rcourse (visited on 23. 06. 2019).

Lehmann, E.L. (1993). The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theo ry or Two? *Journal of the American Statistical Association/ Theory and Methods, 88*(424), 1242–1249. https : / / errorstatistics . files . wordpress . com / 2013 / 11 / lehmann _ 1 -theory-or-2.pdf.

Lei, Rayleigh, Andrew Gelman & Yair Ghitza (2017). The 2008 Election: A Preregistered Replication Analysis. *Statistics and Public Policy, 4*(1), 1–8.

Lenth, Russel V. (8. Mai 2018). *Java Applets for Power and Sample Size* [first published 2006-0!)]. url: https://homepage.divms.uiowa.edu/ ~ rlenth/Power (visited on 22. 06. 2019).

Lenth, Russel V. (2007-07). *Post Hoc Power: Tables and Commentary*. Techn. Ber. https : / / stat.uiowa.edu/sites-/stat.uiowa.edu/files/techrep/tr378.pdf.

Levene, Howard (1960). Robust tests for equality of variances. In Ingram Olkin & Harold Hotelling (Eds.), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. (pp. 278-292). Stanford: Stanford University Press.

Levin Joel, R., & Marascuilo, Leonard A. (1972). Type IV errors and interactions. *Psychological Bulletin 78*(5), 368–374.

Levitt, Heidi M. u. a. (2018). Journal Article Reporting Standards for Qualitative Primary, Quali tative Meta-Analytic, and Mixed Methods Research in Psychology: The APA Publications and Communications Board Task Force Report. *American Psychologist 73*(1), 26–46.

Lienert, Gustav A. & Ulrich Raatz (1998). *Testaufbau & Testanalyse*. Weinheim: Beltz.

Ligges, Uwe (2005). *Programmieren mit R (Statistik und ihre Anwendungen*) (German Edition). Berlin: Springer. ISBN: 9783540207276.

Lindeberg, Jarl Waldemar (1922). Eine neue Herleitung des Exponentialgesetztes in der Wahr scheinlichkeitsrechnung. *Mathematische Zeitschrift, 15,* 211–22. https : / / gdz . sub . uni - goettingen.de/id/PPN266833020_0015

Lindeløv, Jonas Kristoffer (Feb. 2018). *How to compute Bayes factors using lm, lmer, BayesFactor, brms, and JAGS/stan/pymc3*. R-code: https : / / github . com / lindeloev / bayes _ factors. url: https : / / rpubs . com / lindeloev / bayes _ factors (visited on 06. 05. 2019).

Lindley, Dennis V. (1957). A statistical paradox. *Biometrika, 44*, 187–192.

Lindley, Dennis V. (2000). The Philosophy of Statistics. *Journal of the Royal Statistical Society. Series D (The Statistician) 49*(3), 293–337.

Lindley, Dennis V. & Melvin R. Novick (1981). The role of exchangeability in inference. *The Annals of Statistics 9*(1), 45–58. https://projecteuclid.org/euclid.aos/1176345331,

Little, Roderick J.A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*(404), 1198–1202.

Little, Roderick J.A. & Donald B. Rubin (1983). On jointly estimating parameters and missing data by maximizing the complete-data likelihood. *The American Statistician 37*(3), 218– 220.

Little, Roderick J.A. & Donald B. Rubin (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons Inc.

Liu, Charles C. & Murray Aitkin (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology 52*, 362–375.

Locker, Bernard (2009). Doob at Lyon. On his lecture, Application of the Theory of Martingales, at the Lyon Colloquium, June 28 – July 3, 1948. Translated from the French by Ronald Sverdlove. *Electronic Journ@l for History of Probability and Statistic, 5*(1). http://www.jehps.net/juin2009/Locker.pdf.

Loftus, Elisabeth (1998). Falsche Erinnerungen. *Spektrum der Wissenschaft Magazin, 1*, 63. url: https://www.spektrum . de/magazin/falsche-erinnerungen/823559 (visited on 07. 06. 2019).

Loken, Eric & Gelman Andrew (2017). Measurement error and the replication crisis. The assumption that measurement error always reduces effect sizes is false. *Science 355*(6325), 584–585.

Loredo, Tom (1990). From Laplace to Supernova SN 1987A: Bayesian Inference in Astrophysics. In P. F. Fougère (Ed.), *Maximum Entropy and Bayesian Methods*. (Pp 81-142). http://hosting.astro.cornell.edu/staff/loredo/bayes-/L90-LaplaceToSN1987A-scan.pdf. Dordrecht, NL: Kluwer Academic Publishers.

Loredo, Tom (1992). The Promise of Bayesian Inference for Astrophysics. In E. D. Feigelson & G. J. Babu (Eds.), *Statistical Challenges in Modern Astronomy* (pp. 275–297). http://www.astro. cornell.edu/staff/loredo/bayes/promise.pdf. New York: Springer

Lorenz, Hilke (2005). *Kriegskinder. Das Schicksal einer Generation*. BerlList/ Ullstein.

Lortie-Forgues, Hugues & Matthew Inglis (2019). Rigorous Large-Scale Educational RCTs Are Often Uninformative: Should We be Concerned? *Educational Researcher, 48*(3). Supplemental materials: https : / / journals . sagepub . com / doi / suppl / 10 .3102 / 0013189X19832850 / suppl _ file / Lortie - ForguesandInglis(2019)_Supplementary . pdf, Raw data and R-Code: https : / / doi . org / 10 . 6084 / m9 . figshare.c.4421087, S. 158–166. (Visited on 23. 05. 2019).

Loy, Adam (15. Juni 2021). *Diagnostics for mixed/ hierarchical linear models*. https://lib. dr.iastate.edu/etd/13277. Dissertation. Ames, Iowa, USA: Iowa State University.

Loy, Adam, Lendie Follett & Heike Hofman (2015). Variations of Q-Q Plots — the Power of your Eyes! *The American Statistician 70*(2), 202–214. doi: 10.1080/00031305.2015.1077728. arXiv: 1503.02098v1 [stat.ME]. url: http://dx.doi.org/10.1080/00031305.2015.1077728.

Loy, Adam & Heike Hofmann (2013). Diagnostic tools for hierarchical linear models. *WI- REs Comput Stat, 5*, 48–61. doi: 10.1002/wics.1238.

Loy, Adam, Heike Hofmann & Dianne Cook (2016). *Model Choice and Diagnostics for Linear Mixed-Effects Models Using Statistics on Street Corners*. ArXiv e-prints 70(2). http:// arxiv.org/abs/1502.06988v3.

Ludwiger, Illobrand von, (Ed.) (2015). *Ergebnisse aus 40 Jahren UFO-Forschung: Wie die Untersuchungen von MUFON-CES zu einem neuen Weltbild führten*. Rottenburg: Kopp.

lukeprog (29. Aug. 2011). *A history of Bayes' Theorem. LessWrong — a community blog devoted to the art of human rationality*. url: https : / / www . lesswrong . com / posts / RTt59BtFLqQbsSiqd/a-history-of-bayes-theorem.

Lunn, David et al. (3. Okt. 2012). *The BUGS Book*. CRC Press/ Taylor & Francis Inc. ISBN: 1584888490.

Ly, Alexander, Josine Verhagen & Eric-Jan Wagenmakers (2016). Harold Jeffreys's Default Bayes Factor Hypothesis Tests: Explanation, Extension, and Application in Psychology. *Journal of Mathematical Psychology, 72*, 19–32. http://www.alexander- ly.com/wp- content/uploads/2014/09/JeffreysToPTests.pdf.

Mächler, Martin (5. Nov. 2015). *Missing Data Imputation etc: Literature and R packages* [Resources for the Lecture Advanced Topics in Computational Statistics, 2015 | part II]. url: https : / / stat.ethz.ch/ ~ maechler-/adv_topics_compstat/MissingData_Imputation. html.

Magnusson, Kristoffer (29. Nov. 2016). Equivalence, non-inferiority and superiority testing an in- teractive visualization. url: https://rpsychologist.com/d3/equivalence (visited on 29. 06. 2019).

Magnusson, Kristoffer (2018-08-14). *powerlmm: Power Analysis for Longitudinal Multilevel Models* [version 0.4.0] [Vignettes]. Techn. Ber. CRAN. url: https://cran.r- project.org/ web/packages/powerlmm (visited on 26. 06. 2019).

Mahalanobis, P.C. (1936). On the Generalized Distance in Statistics. *Proceedings of National Institute of Siences (India) 2*(1), 49–55.

Marascuilo, Leonard A. & Levin Joel R. (1970). Appropriate Post Hoc Comparisons for Interaction and nested Hypotheses in Analysis of Variance Designs: The Elimination of Type-IV Errors. *American Educational Research Journal 7*(3), 397–421.

Marcelo García, Carlos (1991). *El primer año de enseñanza*. Sevilla: Grupo de Investigación Didáctica de la Universidad de Sevilla.

Marecek, Jeanne (2003). Dancing through minefields: Toward a qualitative stance in psychology. In P. M. Camic, J. E. Rhodes, &L. Yardeley (Eds.), *Qualitative research in psychology: Expanding perspectives in methodology and design* (pp. 49-69). Washington, DC.: American Psychological Association.

Marsman, Maarten et al. (2017). A Bayesian bird's eye view of 'Replications of important results in social psychology'. *Royal Society Open Science 4*. http://www.stat.columbia.edu/ ~ gelman/research/published/birdseye.pdf.

Marx, Karl & Friedrich Engels (1851/ 1964). *Der achtzehnte Brumaire des Louis Bonaparte*. Berlin: Dietz.

Matechou, Eleni (2013-12-04). *Model Checking* [Principles for Statistical Methods — Michaelmas Term 2013]. http :// www . stats . ox . ac . uk / ~ steffen / teaching / principles / ModelChecking.pdf. Oxford/ UK.

Matthews, Robert, Ron Wasserstein & David Spiegelhalter (2017). The ASA's p-value statement, one year on: Significance. *Royal Statistical Association 4*, 38–41. https : / / rss . onlinelibrary.wiley.com/doi/epdf/10.1111/j.1740-9713.2017.01021.x.

Maturana, Humberto R. & Francisco J. Varela (1984). *Der Baum der Erkenntnis. Die biologischen Wurzeln des menschlichen Erkennens*. Bern/ München: Scherz/ Goldmann.

Maxwell, Joseph A. (1996). *A qualitative research design. An interactive approach*. Thousand Oaks: Sage.

Mayo, Deborah G. (1981). In Defense of the Neyman-Pearson Theory of Confidence Intervals. *Philosophy of Science 48*(2), 269–280. https://www.jstor.org/stable/187185?origin = JSTOR-pdf.

Mayo, Deborah G. (2018). *Statistical Inference as Severe Testing. How to get beyond the Statistics Wars*. Cambridge: Cambridge University Press.

Mayo, Deborah G. & Aris Spanos (Apr. 2006). Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction. *The British Journal for the Philosophy of Science 57*(2), 323–357. ISSN: 0007-0882. doi: 10.1093/bjps/axl003. url: http://dx.doi.org/10.1093/bjps/axl003.

Mayring, Philipp (1990). *Einführung in die qualitative Sozialforschung*. München: Psychologische Verlagsunion.

Mayring, Philipp (1995). *Qualitative Inhaltsanalyse, Grundlagen und Techniken*. Weinheim: Deutscher Studien Verlag.

Mayring, Philipp (Feb. 2001). Kombination und Integration qualitativer und quantitativer Ansätze. *Forum Qualitative Sozialforschung, 2*. http://www.qualitative-research.net/fqs-texte/1-01/1-01mayring-d.htm.

McCulloch, C.E. (1987). Tests for equality of variances for paired data. *Communications in Statistics. Theory and Methods 16*, 1377–1391.

McDermott, Robert M. (1985). *Computer-aided logic design*. Indianapolis: Howard W. Sams & Co.

McElreath, Richard (21. Dez. 2015). *Statistical Rethinking*. Apple Academic Press Inc. ISBN: 1482253445.

McGryne, Sharon Bertsch (2011). T*he theory that would not die: How Bayes' rule cracked the enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy*. New Haven & London: Yale University Press.

McShane, Blakeley B. et al. (2019). Abandon statistical significance. *The American Statistican 73*(1), 235–245. http : / / www . stat . columbia . edu / ∼gelman / research / published / abandon.pdf.

Medina Rivilla, Antonio et al. (2002). The methodological complementariness of biograms, in-depth interviews, and discussion groups. In Mechthild Kiegelmann (Ed.), *Qualitative Psychology Nexus: Vol. 2. The role of the researcher in qualitative psychology* (pp. 170–185). Tübingen: Ingeborg Huber Verlag.

Meehl, Paul E. (1967). Theory-testing in psychology and physics: a methodological paradox. *Philosophy of Science, 34,* 103–115. https :// citeseerx . ist . psu . edu / viewdoc / download ? doi = 10.1.1.693.8918&rep = rep1&type = pdf.

Meehl, Paul E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66*, 195–244. http : / / meehl . umn . edu / sites / g / files / pua1696/f/144whysummaries.pdf.

Metropolis, N. (1987). The beginning of the Monte Carlo method. *Los Alamos Science [Special Issue]*, 125–130.

Metropolis, N., A. Rosenbluth et al. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics 21*, 1087–1092.

Metropolis, N. & S. Ulam (1949). The Monte Carlo method. *Journal of the American Statistical Association 44*, 335–341.

Michael, Robert B. u. a. (2013). On the (non)persuasive power of a brain image. Psychono- mic Bulletin and Review 20(4). https://link.springer.com/article/10.3758% 2Fs13423-013-0391-6, S. 720–725.

Michalos, Alex C. (1971). The Popper-Carnap controvery. http : / / www . fitelson . org / confirmation / michalos _ popper _ carnap _ controversy . pdf. The Hague/ Netherlands: Martinus Nijhoff.

Miles, Matthew B & A Michael Huberman (1994). *Qualitative data analysis: An expanded sourcebook*. London: Sage.

Miles, Matthew B. & Michael A. Huberman (1984). *Qualitative data analysis. A sourcebook of new methods*. Beverly Hills: Sage.

Miller, Evan (2015). *Formulas for Bayesian A/B testing*. url: https://www.evanmiller.org/bayesian-ab-testing.html (visited on 01. 10. 2015).

Milligan, G.W. & M.C. Cooper (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika 50*, 159–179.

Minka, Thomas (1998). *Pathologies of Orthodox Statistics* [last modified 2004-I2-I0]. https://tminka.github.io/papers-/minka-pathologies.pdf. MIT Media Lab note.

Mlinaric, Ana, Martina Horvat & Vesna Šupak Smolcic (2017). Dealing with the positive publication bias: Why you should really publish your negative results. *Biochem Med (Zagreb) [online] 27*(3): 030201. https : / / www . ncbi . nlm . nih . gov / pmc / articles / PMC5696751/pdf/bm-27-3-030201.pdf.

Mohammad-Djafari, Ali (2012-11-12). *Yet Another Analysis of Dice Problems* [Presented at MaxEnt2002, the 22nd International Workshop on Bayesian and Maximum Entropy methods (Aug. 3-9, 2002, Moscow, Idaho, USA). To appear in Proceedings of American Institute of Physics]. url: https://arxiv.org/abs/physics/0211049 (visited on 05. 06. 2019).

Mojena, R. (1977). Hierarchical grouping methods and stopping rules: an evaluation. *Computer Journal 20*, 359–363.

Morey, Richard (12. Feb. 2014). *Bayes factor t tests, part I*. url: http : / / bayesfactor . blogspot.com/2014/02/bayes-–-factor-t-tests-part-1.html.

Morey, Richard D., Rink Hoekstra et al. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin and Review 23*, 103–123. https://www.ncbi.nlm.nih.gov/ pmc/articles/PMC4742505/pdf-/13423_2-015_Article_947.pdf.

Morey, Richard D. & Jeffrey N. Rouder (Dez. 2011). Bayes Factor Approaches for Testing Interval Null Hypotheses. *Psychological Methods 16*(4), 406–419. https : / / www . researchgate . net / publication / 51519224 _ Bayes _ Factor Approaches _ for _ Testing _ Interval_Null_Hypotheses

Morey, Richard D. & Eric–Jan Wagenmakers (2014). Simple relation between one-sided and two-sided Bayesian point-null hypothesis tests. *Statistics and Probability Letters 92, 121-124*. http://www.ejwagenmakers.com/2014-/MoreyWagenmakers2014.pdf.

Morgan, W.A. (1939). A test for the significance of the difference between the two variances in a sample from a normal bivariate population. *Biometrika 31*, 13–19.

Morse, Janice M. (2003). Principles of mixed methods and multimethod research design. In Abbas Tashakkori & Charles Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 189–208). Hrsg. von. Thousand Oaks: Sage Publ.

Mosteller, F. (1948). A k-sample slippage test for an extreme population. *The Annals of Mathematical Statististics 19*, 58–65.

Mosteller, F. & John Wilder Tukey (1977). *Data Analysis and Regression: A Second Course in Statistics*. Reading, Massachusetts: Addison-Wesley.

Mullican, Timothy (22. Juni 2021). *Calculating Pi: My attempt at breaking the Pi World Record* (last update 2020-0I-2!). url: https :// blog . timothymullican . com / calculating - pi-my-attempt-breaking-pi-record (visited on 26. 06. 2019).

Munroe, Randall (o. D.). *P-Values. xkcd.com — A webcomic of romance, sarcasm, math, and language.* url: https://xkcd.com/1478.

Murad, Mohammad Hassan et al. (2018). The effect of publication bias magnitude and direction on the certainty in evidence. *BMJ Evidence-Based Medicine 23*(3), 84–86. https://ebm.bmj. com/content/23/3/84. doi: 10.1136/bmjebm-2018-110891.

Murphy, Kevin P. (2007). *Conjugate Bayesian analysis of the Gaussian distribution* [latest version: 200J-I0-03]. Techn. Ber. University of British Columbia. url: https://www.cs.ubc.ca/ ~ murphyk/Papers/bayesGauss.pdf (visited on 03. 06. 2019).

Murre, Jaap M.J. & Joeri Dros (2015). Replication and analysis of Ebbinghaus' forgetting curve. *PLoS One 10*(7). https : / / www . ncbi . nlm . nih . gov / pmc / articles / PMC4492928.

Muth, Chelsea, Zita Oravecz & Jonah Gabry (2018). User-friendly Bayesian regression modeling: A tutorial with rstanarm and shinystan. *The Quantitative Methods for Psychology[TQMP], 14*(2), 99-119. https://www.tqmp.org-/RegularArticles/vol14-2/p099/p099.pdf

Nadarajah, Saralees & Samuel Kotz (2007). Statistical distribution of the difference of two proportions. *Statistics in Medicine 26*(18), 3518–3523.

Nalborczyk, Ladislas (23. Jan. 2018). *Checking the asumption of independence in binomial trials using posterior predictive checking.* url: https://www.barelysignificant.com/post/ ppc/ (visited on 03. 06. 2019).

National Geographic (12. Apr. 2012). *Titanic: A remembrance. Documentary movie about how the Titanic sank and broke apart into pieces* [6:25 min], New CGI of How Titanic Sank | Ti- tanic 100 [2:41 min]. National Geographic. https : / / www . youtube . com / watch ? v = 1PhMWUoPDsk, https://www.youtube.com/watch?v = FSGeskFzE0s.

Nau, Robert F. (2001). De Finetti was right: Probability does not exist. *Theory and Decision 51*, 89–124. https://faculty.fuqua.duke.edu/ ~ rnau/definettiwasright.pdf.

Neal, Radford M. (2011). MCMC Using Hamiltonian Dynamics. In: Steve Brooks et al. *(*Eds.*), Handbook of Markov Chain Monte Carlo* (pp. 113–162). London: Chapman & Hall/ CRC, S.. url: http :// www . mcmchandbook . net / HandbookChapter5 . pdf (visited on 06. 05. 2019).

Neto, João (Juli 2014). *BUGS tutorial* (by example). url: http://www.di.fc.ul.pt/ ~ jpn/ r/bugs/bugs_tutorial.html (visited on 07/2014).

Neto, João (Jan. 2015). *Bayesian Decision Theory*. url: http://www.di.fc.ul.pt/ ~ jpn/r/ decision/index.html (visited on 28. 05. 2019).

Neyman, Jerzy (1935). On the Problem of Confidence Intervals. *The Annals of Mathematical Statistics* 6(3), 111–116. https :// projecteuclid . org / download / pdf _1/ euclid . aoms /1177732585

Neyman, Jerzy (1937). Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences 236*, 333–380. https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.1937.0005.

Neyman, Jerzy (Ed.) (1950). *First Course in Probability and Statistics*. New York: Henry Hold & Co.

Neyman, Jerzy (1955). The Problem of Inductive Inference. *Communications on Pure and Applied Mathematics VIII*, 13–46.

Neyman, Jerzy & Egon Pearson (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231, 289–337.

NIST/SEMATECH (2003). *e-Handbook of Statistical Methods* (Engineering Statistics Handbook) [2003-0I-06, latest version 20I2-06-28]. url: http : / / www . itl . nist . gov / div898 / handbook.

Noether, Gottfried E. (Ed.) (1967). *Elements of Nonparametric Statistics*. New York: John Wiley & Sons.

Nosek, Brian A. & Daniël Lakens (2014). A method to increase the credibility of published results. *Perspectives on Psychological Science 45*(3), 137–141. https : / / econtent . hogrefe . com/doi/pdf/10.1027/1864-9335/a000192.

Nosek, Brian A., Jeffrey R. Spies & Matt Motyl (2012). Scientific Utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science 7*(6), 615–631. url: https://arxiv.org-/pdf/1205.4251.pdf.

Nowak, Martin A. (2006). Five rules for the evolution of cooperation. *Science 314*(5805), 1560–1563. https : / / www . ncbi . nlm . nih . gov / pmc / articles / PMC3279745 / pdf / nihms49939.pdf.

Nyanaponika Thera & Hellmut Hecker (Eds.) (1997). *Great Disciples of the Buddha. Their lives, their works,. their legacy*. Boston/ Kandy, Sri Lanka: Wisdom Publications/ Buddhist Publication Society.

O'Hagan, Anthony et al. (2006). *Uncertain judgements. Eliciting expert's probabilities*. Chichester/ UK: John Wiley & Sons.

Oelker, Laura & Sybille Klormann (2012-11-06). *Wahlsystem in den USA. Wie wird der US-Präsident gewählt* [latest version: 2017-01-23]. Zeit Online. url: https://www.zeit. de/politik/ausland/2012-11/usa-wahl-wahlsystem (visited on 07. 06. 2019).

Oevermann, Ulrich (1979a). Die Methodologie einer objektiven Hermeneutikund ihre allgemeine forschungslogische Bedeutung in den Sozialwissenschaften In Allert, Tilman, Konau, Elisabeth, & Krambeck, Jürgen (1979), *Interpretative Verfahren in den Sozial- und Textwissenschaften* (pp. 352–434). Stuttgart: JB Metzlersche Verlagsbuchhandlung.

Oevermann, Ulrich (1979b). Ansätze zu einer soziologischen Sozialisationstheorie. *Sonderheft 2I der Kölner Zeitschrift für Soziologie und Sozialpsychologie*, pp. 143–168. Hrsg. von M.R. Lepsius. Opladen: Westdeutscher Verlag.

Oevermann, Ulrich (1981). *Fallrekonstruktionen und Strukturgeneralisierung als Beitrag der objektiven Hermeneutik zur soziologisch-strukturtheoretischen Analyse*. Unveröffentlichtes Manuskript, Universität Frankfurt, http://user.uni - frankfurt.de/ ~ hermeneu/Fallrekonstruktion-1981.rtf.

Oevermann, Ulrich (1993). Die objektive Hermeneutik als unverzichtbare methodologische Grundlage für die Analyse von Subjektivität. Zugleich eine Kritik der Tiefenhermeneutik. In T. Jung & S. Müller-Doohm (Eds.), *»Wirklichkeit« im Deutungsprozeß. Verstehen und Methoden in den Kultur- und Sozialwissenschaften* (pp. 106–189). Frankfurt am Main: Suhrkamp.

Oevermann, Ulrich (1996a). *Konzeptualisierung von Anwendungsmöglichkeiten und praktischen Arbeitsfeldern der objektiven Hermeneutik. Manifest der objektiv hermeneutischen Sozialforschung*. Unveröffentlichtes Manuskript, Universität Frankfurt am Main.

Oevermann, Ulrich (1996b). Theoretische Skizze einer revidierten Theorie professionalisierten Handelns. In A. Combe & W. Helsper (Eds.), *Pädagogische Professionalität: Untersuchungen zum Typus pädagogischen Handelns* (pp. 70–182). Frankfurt am Main: Suhrkamp.

Oevermann, Ulrich (1997). *Gebildeter Fundamentalismus oder pragmatische Krisenbewältigung*. Unveröffentlichtes Manuskript, korrigierte Version, Universität Frankfurt am Main.

Oevermann, Ulrich (1998a). *Der Fokus stationäre Suchttherapie aus der Sicht einer Theorie professionalisierten Handelns*. Vortragsmitschrift, Start Again, Zürich.

Oevermann, Ulrich (1998b). *Projektskizze, Teil I: »Struktur und Genese professionalisierter Praxis als Ort der Krisenbewältigung«, Teil II: »Bewährungsdynamik, lebenspraktische Krisenbewältigung und die Entstehung von Habitusformationen und Deutungsmustern des Rationalisierungsprozesses«*. Unveröffentlichtes Manuskript, Universität Frankfurt am Main.

Oevermann, Ulrich (2000). Die Methode der Fallrekonstruktion in der Grundlagenforschung sowie der klinischen und pädagogischen Praxis. In Klaus Kraimer (Ed.), *Die Fallrekonstruktion. Sinnverstehen in der sozialwissenschaftlichen Forschung* (pp. 58–156). Frankfurt am Main:Suhrkamp.

Oevermann, Ulrich (2002). *Klinische Soziologie auf der Basis der Methodologie der objektiven Hermeneutik — Manifest der objektiv hermeneutischen Sozialforschung*. Unveröffentlichtes Manuskript, korrigierte Version, Universität Frankfurt am Main, http://www.ihsk.de/ManifestWord.doc.

Oevermann, Ulrich (2004). Sozialisation als Prozeß der Krisenbewältigung. In Dieter Geulen & Hermann Veith (Eds.), *Sozialisationstheorie interdisziplinär — aktuelle Perspektiven* (pp. 155–181). Stuttgart: Lucius und Lucius.

Oevermann, Ulrich, Tilman Allert & Elisabeth Konau (1980). Zur Logik der Interpretation von Interviewtexten. Fallanalyse anhand eines Interviews mit einer Fernstudentin. In Th. Heinze, H.W. Klusemann & Hans-Georg Soeffner, *Interpretationen einer Bildungsgeschichte*, (pp. 15–69).Weinheim: Beltz

Oevermann, Ulrich, Tilman Allert, Elisabeth Konau & Jürgen Krambeck (1979). Die Methodologie einer "objektiven Hermeneutik" und ihre allgemeine forschungslogische Bedeutung in den Sozialwissenschaften. In Th. Heinze, H.W. Klusemann & Hans-Georg Soeffner, *Interpretative Verfahren in den Sozial- und Textwissenschaften (*pp. 352–434*).* Stuttgart: Metzler.

Oganisian, Arman (7. Aug. 2007). *Bayesian Simple Linear Regression with Gibbs Sampling in R*. R-Code: https://github.com/stablemarkets/BayesianTutorials. url: https: // stablemarkets . wordpress . com / 2017 / 08 / 07 / bayesian - simple - linear - regression-with-gibbs-sampling-in-r (visited on 05. 06. 2019).

Oldenbürger, Hartmut A. (1981). *Methodenheuristische Überlegungen und Untersuchungen zur »Erhebung« und Repräsentation kognitiver Strukturen*. Dissertationzur Erlangung des Doktorgrades der Mathematisch-Naturwissenschaftlichen Fachbereiche der Georg-August-Universität Göttingen. Göttingen/ Braunschweig: Technische Universität Carolo-Wilhelmina.

Oldenbürger, Hartmut A. (1994). Clusteranalyse. In Theo Herrmann (Ed.), *Enzyklopädie der Psychologie. Bd. 4, Kap. 7*, 390–439. Göttingen: Hogrefe.

Oliphant, Travis E. (2006-12-05). *A Bayesian perspective on estimating mean, variance, and standard-deviation from data*. Techn. Ber. Faculty Publications, Brigham Young University, BYU Scholars Archive. url: http : / / hdl . lib . byu . edu / 1877 / 438 (visited on 07. 01. 2020).

Onwuegbuzie, Anthony J. & Larry G. Daniel (2003). Typology of Analytical and Interpretational Errors in Quantitative and Qualitative Educational Research. *Current Issues in Edu- cation [Online]* 6(2). http://cie.asu.edu/ojs/index.php-/cieatasu/article/view/1609.

Orwell, George (1946-01-12). *A Nice Cup of Tea* [Reprinted: The Collected Essays, Journalism and Letters of George Orwell, 1968]. Evening Standard. url: http :// orwell . ru / library/articles/tea/english/e_tea (visited on 05. 06. 2019).

Ovens, Matthew (2012). Cohen (1994). *The Earth is round (p < .05)* [last update 2018-06-I9)]. YourStats-Guru. url: https://www.yourstatsguru.com/epar/rp- reviewed/cohen1994/?v = 4442e4af0916 (visited on 29. 05. 2019).

Pearl, Judea (2009). *Causality: Models, Reasoning and Inference*. New York: Cambridge University Press.

Pearl, Judea (1999-04). *Simpson's Paradox: An Anatomy*. Techn. Ber. http://bayes.cs.ucla.edu/R264.pdf. Cognitive Systems Laboratory. Computer Science Department. University of California. Los Angeles. CA 90024 [UCLA].

Peirce, Charles S. (1965). *Collected papers of Charles Sanders Peirce: Vol V.: Pragmatism and pragmaticism; Vol. VI: Scientific metaphysics* (C. Harsthorne, and P. Weiss, Eds.) Cambridge, MA: The Belknap Press of Harvard University Press.

Penfield, Paul (2003). *Principle of Maximum Entropy* [version I.0.2] [course notes: 6.050J / 2.II0J, In- formation and Entropy, Spring 2003]. Chapter 9 and 10: https :// mtlsites . mit . edu / Courses /6. 050 / 2003 / notes / chapter9 . pdf, https :// mtlsites .mit . edu / Courses/6.050/2003/notes/chapter10.pdf.

Perry, Gina (2013). *Behind the Shock Machine: The Untold Story of the Notorious Milgram Psychology Experiments*. New York: The New Press.

Peters, Ole & Alexander Adamou (10. Juni 2015). *The evolutionary advantage of cooperation*. Unpublished manuscript. url: https://arxiv.org/abs/1506.03414 (visited on 24. 05. 2018).

Pham-Gia, T., N. Turkkan & P. Eng (Jan. 1993). Bayesian analysis of the difference of two proportions. *Communications in Statistics - Theory and Methods 22*(6), 1755–1771. ISSN: 1532- 415X. doi: 10 . 1080 / 03610929308831114. url: http :// dx . doi . org / 10 . 1080 /03610929308831114.

Pinheiro, Jose C. & Bates, Douglas M. (11. Apr. 2009). *Mixed-Effects Models in S and S-Plus*. Springer Verlag GmbH. ISBN: 1441903178.

Pitman, E.J.G. (1939). A note on normal correlation. *Biometrika 31*, 9–12.

Platon (1994). *Politeia*. Platon. Sämtliche Werke. Band 2, pp. 195–538. Hrsg. von Burghard König. Reinbek b. Hamburg: Rowohls Enzyklopädie.

Plummer, Martyn (2017-06-28). *JAGS Version 4.3.0 user manual*. url: https://sourceforge. net / projects / mcmc - jags / files / Manuals / 4 . x / jags _ user _ manual . pdf / download.

Pólya, George (1920). Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem. *Mathematische Zeitschrift, 8*(3–4), S. 171–181. https://gdz.sub.uni- goettingen.de/id/PPN266833020_0008.

Pólya, George (1954a). *Mathematics and Plausible Reasoning. Volume I: Induction and Analogy in Mathematics*. (Reprint 1990). New Jersey: Princeton University Press.

Pólya, George (1954b). *Mathematics and Plausible Reasoning. Volume II: Patterns of Plausible Inference.* (Reprint 1990). New Jersey: Princeton University Press.

Popper, Karl (1943). *Logik der Forschung. II. Auflage 2005*. Herausgegeben von Herberth Keupp, Tübingen: Mohr..

Population Research [OPR], Office of (2019). *Switzerland Socio-Economic Variables*, I870-1930. url: https://opr.princeton.edu/archive/pefp/switz.aspx (visited on 20. 05. 2019).

Porter, M. F. (1980). An algorithm for suffix stripping. *Program 14*(3), 130–137. Algorithm and Implementations in: https://tartarus.org/martin/PorterStemmer, R Code: https://tartarus.org/martin/PorterStemmer/R.txt.

Presidential Debates (CPD), Commission on (1987). *CPD's primary purpose is to sponsor and produce the quadrennial general election debates and to undertake research and educational activities relating to the debates. The organization, which is a nonprofit, nonpartisan, 501(c)(3) corporation, sponsored all of the presidential debates in 1988, 1992, 1996, 2000, 2004, 2008, 2012, and 2016.* url: https://debates.org/about-cpd (visited on 05. 06. 2019).

Project, Many Labs Replication (o. D.). *Investigating Variation in Replicability: A Many LabsReplication Project*. OSF — Open Science Framework/ COS — Center for Open Science. url: https://osf.io/wx7ck.

Pulskamp, Richard J. (21. Aug. 2019). *Pierre Simon Laplace on Probablity and Statistics* [Sources in the history of probability and statistics]. url: https : / / web . archive . org / web / 20190821194930 / http :// cerebro . xu . edu / math / Sources / Laplace / index . html (visited on 21. 08. 2019).

Quenouille, M. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society, Soc. Series B 11*, 18–84.

R Development Core Team (18. Juni 2018). *CRAN Task View: Robust Statistical Methods*. https://cran.r-project.org/web/views/Robust.html. Vienna/ Austria.

R Development Core Team (8. Mai 2019a). CRAN Task View: Bayesian Inference. https : / / cran.r-project.org/web/views-/Bayesian.html. Vienna/ Austria.

R Development Core Team (8. Mai 2019b). *CRAN Task View: Cluster Analysis and Finite Mixture Models*. https : / / cran . r - project . org / web / views / Cluster . html. Vienna/Austria.

R Development Core Team (11. Mai 2019c). *CRAN Task View: Missing Data*. https://cran.r- project.org/web/views-/MissingData.html. Vienna/Austria.

R Development Core Team (2019d). *R: A language and environment for statistical computing*. https://www.r-project.org. Vienna/ Austria.

R Development Core Team (1. Nov. 2021). R Data Import/ Export. https : / / cran . r - project.org/doc/manuals/r-release-/R-data.html. Vienna/ Austria.

Radebold, Hartmut, Bohleber, Werner & Zinnecker, Jürgen (Eds.) (2008). *Transgenerationale Weitergabe kriegsbelastender Kindheiten. Interdisziplinäre Studien zur Nachhaltigkeit historischer Erfahrungen über vier Generationen*. Weinheim/ München: Juventa.

Radin, D.I. (1997). Unconscious perception of future emotions: An experiment in presentiment. *Journal of Scientific Exploration 11*, 163–180.

Radin, D.I. (2006). *Entangled minds: Extrasensory experiences in a quantum reality*. New York: Paraview Pocket Books.

Raftery, Adrian E. (1999). Bayes Factors and BIC. Comment on A Critique of the BayesianIn- formation Criterion for Model Selection. *Sociological Methods and Research 22*(3), 411–427. url: https : / / sites . stat . washington . edu / raftery / Research / PDF /weakliem1999.pdf.

Raftery, Adrian E. & Lewis, S.M. (1992). One long run with diagnostics: Implementation strategies for Markov Chain Monte Carlo. *Statistical Science 7*, 493–497.

Raftery, Adrian E. & Lewis, S.M. (1995). The number of iterations, convergence diagnostics and generic Metropolis algorithms. In: W.R. Gilks, D.J. Spiegelhalter & S. Richardson (Eds.), *Practical Markov Chain Monte Carlo*. London: Chapman & Hall.

Ragin, Charles C. (1987). *The comparative method. Moving beyond qualitative and quantitative strategies*. Berkeley: Berkeley University Press.

Ragin, Charles C. (2000). *Fuzzy-Set Social Science*. Berkeley: Berkeley University Press.

Raineri, Emanuele, Dabad, Marc & Heath, Simon (2014-05-13). A Note on Exact Differences between Beta Distributions in Genomic (Methylation) Studies. *PLoS One 9*(5): e97349. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4019606.

Reich, Wilhelm (1933). *Massenpsychologie des Faschismus. Zur Sexualökonomie der politischen Reaktion und zur proletarischen Sexualpolitik*. https://archive.org/details/Reich_1933_Massenpsychologie_k. Kopenhagen: Verlag für Sexualpolitik.

Reichertz, Jo (2000). Abduktion, Deduktion und Induktion in der qualitativen Forschung. In: Uwe Flick, Ernst v. Kardorff & Ines Steinke (Eds.), Qualitative Sozialforschung. Ein Handbuch (pp. 276–286). Reinbek bei Hamburg: Rowohlts Enzyklopädie.

Reichertz, Jo (2002). Prämissen einer hemerneutisch wissenssoziologischen Polizeiforschung. *Forum Qualitative Sozialforschung 3* (1). http://www.qualitative- research.net/ index.php/fqs/article/view/881/1920.

Reimer, Torsten & Jörg Rieskamp (2007). Fast and frugal heuristics. Encylopedia of Social Psychology. In: R. Baumeister, & K.D. Vohs (Eds.), pp. 347–349. Sage Publications. url: https://www.researchgate.net/publication-/228509269_Fast_and_frugal_heuristics.

Rheem, Hansol (11. Apr. 2017). *The most famous tea party in the history of statistics: The comparison of Frequentismand Bayesianism*. url: https://medium.com/humansystemsdata/the-most-famous-tea-party-in-the--history-of-statistics-the-comparison- of-frequentism-and-3994bc16037b (visited on 05. 06. 2019).

Richardson, John T.E. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review 6*(2), 135–147. Pre-print: https :// psyarxiv . com/b7z4q.

Rickert, Joseph (30. Nov. 2016). *Missing Values, Data Science and R*. R Views — An R community blog edited by R Studio, Boston MA. url: https://rviews.rstudio.com/2016/11/ 30/missing-values-data-science-and-r (visited on 23. 06. 2019).

Rietz, Christian, Georg Rudinger & Johannes Andres (1996). Lineare Strukturgleichungsmodelle. In: Edgar Erdfelder et al. (Eds.), Handbuch Quantitative Methoden. (Kap. I, pp. 253–268.) Weinheim: Beltz: PVU.

Rinker, Tyler (2021). *qdap Package Vignette*. url: https :// trinker . github . io / qdap / vignettes/qdap_vignette.html (visited on 24. 11. 2021).

Ripley, Ruth u. a. (2019-05-21). *Manual for RSiena*. Manual. University of Oxford: Department of Statistics; Nuffield College | University of Groningen: Department of Sociology. url: http:// www . stats . ox . ac . uk / ~snijders / siena / RSiena _ Manual . pdf (visited on 23. 06. 2019).

Ritchie, Stuart J., Richard Wiseman & Christopher C. French (2012). Failing the Future: Three Unsuccessful Attempts to Replicate Bem's 'Retroactive Facilitation of Recall' Effect. *PLoS ONE [Open Access Online] 7*(3). https://journals.plos.org/plosone/article/ file?id = 10.1371/journal.pone.0033423&type = printable.

Robbins, Herbert (1956). *An Empirical Bayes Approach to Statistics*. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume I: Contributions to the Theory of Statistics (pp. 157-163). Hrsg. von University of California. Berkely/ California: University of California Press. url: https://projecteuclid.org/euclid.bsmsp /1200501653.

Robert, Christian P. (2013). *On the Jeffreys-Lindley's Paradox*. Working Papers 2013-46. https :// EconPapers . repec . org / RePEc : crs : wpaper : 2013 - 46. Center for Research in Economics and Statistics.

Robinson, Andrew (26. Apr. 2008). *[R-sig-ME] interpreting significance from lmer results for dummies (like me)*. url: https://stat.ethz.ch/pipermail/r- sig- mixed- models/ 2008q2/000904.html.

Robinson, Andrew P., Remko Duursma & John Marshall (2005). A Regression-Based Equivalence Test for Model Validation: Shifting the Burden of Proof. *Tree Physiology 25*(7), 903–913. url: https : / / www . researchgate . net / publication / 7868456 _ A Regression - Based _ Equivalence _ Test _ for _ Model _ Validation _ Shifting _ the_Burden_of_Proof (visited on 29. 06. 2019).

Robinson, Daniel H. & Howard Wainer (Dez. 2001). *On the past and future of Null Hypothesis Significance Testing*. Techn. Ber. Princeton, NJ 08541. url: https :// www . ets . org / Media/Research/pdf/RR-01-24-Wainer.pdf.

Robinson, Terry E. & Kent C. Berridge (1995). The mind of an addicted brain. Neural sensitization of wanting versus liking. *Current Directions in Psychological Science 4* (3), 71–76.
Rocco, Tonette S. (2010). Criteria for evaluating qualitative studies. *Human Reources Development International 13*(4), 375–378. https : / / www . researchgate . net / publication / 248996833_Criteria_for_evaluating_qualitative_studies.
Rogozhnikov, Alex (19. Dez. 2016). *Hamiltonian Monte Carlo explained. Brillintly wrong — thoughts on science and programming* [Blog]. url: https://arogozhnikov.github. io/2016/12/19/markov_chain_monte_carlo.html (visited on 05. 06. 2019).
Rohwer, Götz (2011). Qualitative Comparative Analysis. A Discussion of Interpretations. *European Sociological Review 27*, 728–740. http :// www . stat . ruhr - uni - bochum . de / papers / dqca.pdf.
Romeijn, Johannes, Richard Morey & J.N. Rouder (2016). The philosophy of Bayes' factors and the quantification of statistical evidence. *Journal of Mathematical Psychology, 72*, 6–18. https : / / www . rug . nl / research / portal / files / 42474996 / 1 _ s2 . 0 _S0022249615000723_main.pdf.
Rose, Evangeline M. et al. (2018). A new statistical method to test equivalence: an application in male and female eastern bluebird song. *Animal Behaviour 145*, 77–85.
Rost, Jürgen (2004). *Lehrbuch Testtheorie / Testkonstruktion*. Huber Hans. ISBN: 3456839642.
Rost, Jürgen (2003). Zeitgeist und Moden empirischer Analysemethoden [45 Absätze]. *Forum Qualitative Sozialforschung 4*(2) Art. 5. http://nbn-resolving.de/urn:nbn:de: 0114-fqs030258.
Rouder, Jeffrey N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin and Review, 21*, 301–308. http://pcl.missouri.edu/sites/default/files/Rouder- PBR-2014.pdf.
Rouder, Jeffrey N., Jun Lu et al. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin and Review 12*(2), 195–223. https : / / link . springer . com / content/pdf/10.3758/BF03257252.pdf.
Rouder, Jeffrey N. & Richard D. Morey (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin and Review 18*, 682–689. http :// pcl . missouri . edu / sites / default/files/Rouder.Morey_.2011.pbr_.pdf.
Rouder, Jeffrey N. & Richard D. Morey (2012). Default Bayes Factors for Model Selection in Regression. *Multivariate Behavioral Research 47*(6), 877–903. http://pcl.missouri.edu/ sites/default/files/Rouder.Morey_.MBR_.2012.pdf.
Rouder, Jeffrey N., Richard D. Morey & Jordan M. Prvince (2013). Comment — A Bayes Factor Meta-Analysis of Recent Extrasensory Perception Experiments: Comment on Storm, Tressoldi, and Di Risio (2010). *Psychonomic Bulletin 139*(1), 241–247. http://pcl.missouri.edu/sites/default/files/RouderEtAl2013PsychBull.pdf.
Rouder, Jeffrey N., Richard D. Morey, Paul L. Speckman u. a. (2012). Default Bayesfactors for ANOVA designs. *Journal of Mathematical Psychology 56*, 356–374. http :// pcl . missouri . edu/sites/default/files/Rouder.JMP_.2012.pdf.
Rouder, Jeffrey N., Paul L. Speckman u. a. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review 16*(2), 225–237. http://pcl.missouri.edu/ sites/default/files/Rouder.bf_.pdf.
Rousseeuw, P. & Yohai, V. (1984). Robust Regression by Means of S-Estimators. In J. Franke, W. Härdle & D. Martin (Eds.), *Robust and Nonlinear Time Series Analysis. Lecture Notes in Statistics, Vol. 26* (pp. 256–272). New York: Springer.
Rowlinson, J.S. (1970). Probability, Information, and Entropy. *Nature 225*, 1196–1198.
Rubin, Donald B. (1981). The Bayesian Boostrap. *The Annals of Statistics 9*(1), 130–134.
Rubin, Donald B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics 12*(4), 1151–1172. https : / / projecteuclid . org /euclid.aos/1176346785.
Rubin, Donald B., Hrsg. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.
Rubin, Donald B. (1996). Multiple Imputation after 18+ Years. *Journal of the American Statistical Association 434*, 473–489.
Rudolf, G. u. a. (2001). Strukturelle Veränderungen in psychoanalytischen Behandlungen — Zur Praxisstudie analytischer Langzeittherapien (PAL). In: U. Stuhr, M. Leuzinger-Bohleber & M. Beutel (Eds.), Langzeitpsychotherapie. Perspektive für Therapeuten und Wissenschaftler. Stuttgart: Kohlhammer.
Russell, Bertrand (1908). Mathematical logic as based on theory of types. *American Journal of Mathematics, 30*, 222–262.
Ryu, Choonghyun (2019-03-16). *Exploratory Data Analysis [Vignette]. R-package dlookr: Tools for Data Diagnosis, Exploration, Transformation* [version 0.3.9)]. Techn. Ber. https://cran.r- project.org/web/packages/dlookr. CRAN.
Saldaña, Johnny (2009). *An introduction to codes and coding The coding manual for qualitative researchers*. Thousand Oaks: Sage.
Saldaña, Johnny (2015). *The Coding Manual for Qualitative Researchers* (3rd ed.) Thousand Oaks: Sage.
Salsburg, David (2001). *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century: How Statisticians Revolutionized Science in the 20th Century*. New York: W.H.Freeman & Co Ltd.
Sánchez-Espigares, José A. & Jordi Ocaña (2009-07). *An R implementation of bootstrap procedures for mixed models*. https : / / www . r - project . org / nosvn / conferences / useR - 2009 / slides / SanchezEspigares + Ocana . pdf. Talk at the The R User Conference 2009. July 8-10. Agrocampus-Ouest. Rennes. France.
Sanderson, Grant (8. Okt. 2019). *Why do prime numbers make these spirals?* url: https://www. youtube.com-/watch?v=EK32jo7i5LQ (visited on 29. 07. 2020).
Sandvik, L. & B. Olsson (1982). A nearly distribution-free test for comparing dispersion in paired samples. *Biometrika, 69*, 484–485.
Sartre, Jean Paul (1964). *Marximus und Existentialisums. Versuch einer Methodik*. Reinbek bei Hamburg: Rowohlt.

Sayadaw, Pa-Auk Tawya (1999/2019). *Knowing and Seeing [Fifth Revised Edition]. Talks and Questions and Answers at a Meditation Retreat in Taiwan*. 15 Teo Kim Eng Road, Singapore 416385: Pa-Auk Meditation Centre, Singapore 2019: A gift in the public domain, the material cannot be copyrighted. url: https :// drive . google . com / file /d/ 1qwl - bqy180Foo5kT7 _ ABpp3uPiv0hEdi/view (visited on 29. 05. 2019).

Sayadaw, Pa-Auk Tawya (2009/2012). *The Workings of Kamma* [Second Revised Edition]. 15 Teo Kim Eng Road, Singapore 416385: Pa-Auk Meditation Centre (Singapore). 2009/ 2012. PDF by PAMC (Singapore) 03/2013. url: https : / / drive . google . com / file / d /1tAhYl7YCCv03qRKCFnmlRwwmdbvFkr1u/view (visited on 29. 05. 2019).

Sayre, S. (2001). *Qualitative methods for marketplace research*. Thousand Oaks: Sage.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Scheele, Brigitte (1988). Rekonstruktionsadäquanz: Dialog-Hermeneutik. In: Norbert Groeben et al. (Eds.), *Das Forschungsprogramm Subjektive Theorien. Eine Einführung in die Psychologie des reflexiven Subjekts* (pp. 126-179). Tübingen: Francke.

Scheele, Brigitte & Norbert Groeben (1988). *Dialog-Konsens Methoden zur Rekonstruktion Subjektiver Theorien*. Tübingen: Francke.

Schimack, Ulrich (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods 17*(4), 551–566.

Schimmack, Ulrich (30. Dez. 2014). *The Test of InsujJicient Variance (TIVA): A New Tool for the Detection of Questionable Research Practices*. url: https : / / replicationindex . com / 2014/12/30/the-test-of-insufficient- -variance-tiva-a-new-tool-for- the-detection-of-questionable-research-practices (visited on 29. 05. 2019).

Schimmack, Ulrich (24. März 2015a). *An Introduction to Observed Power based on Yuan and Maxwell (2005)*. url: https : / / replicationindex . com / 2015 / 03 / 24 / an - introduction- to- observed- power- based- on- yuan- and- maxwell- 2005 (visited on 29. 05. 2019).

Schimmack, Ulrich (30. Apr. 2015b). *Replacing p-values with Bayes-Factors: A Miracle Cure for the Replicability Crisis in Psychological Science*. url: https : / / replicationindex . com / 2015 / 04 / 30 / replacing - p - values - with - bayes - factors - a - miracle - cure- for- the- replicability- crisis- in- psychological- science (visited on 29. 05. 2019).

Schimmack, Ulrich (9. Mai 2015c). *Why Psychologists Should Not Change The Way They Analyze Their Data: The Devil is in the Default Prior*. url: https://replicationindex.com/ 2015 / 05 / 09 / why - psychologists - should - not - change - the - way - they - analyze - their - data - the - devil - is - in - the - default - piror (visited on 29. 05. 2019).

Schimmack, Ulrich (31. Jan. 2016a). *A Revised Introduction to the R-Index* [The Replicability-Index: Quantifying Statistical Research Integrity]. url: https : / / replicationindex . com / 2016/01/31/a-revised-introduction-to-the-r-index (visited on 29. 05. 2019).

Schimmack, Ulrich (4. März 2016b). *My journey towards estimation of replicability of psychological research*. url: https :// replicationindex . com / 2016 / 03 / 04 / my - journey - towards - estimation - of- replicability - of - psychological - research (be- sucht am 29. 05. 2019).

Schimmack, Ulrich (14. Jan. 2016c). *On the Definition of Statistical Power*. url: https : / / replicationindex.com/2016/01- /14/on- the- definition- of- statistical- power.

Schimmack, Ulrich (30. Juni 2016d). *Wagenmakers' Default Prior is Inconsistent with the Observed Results in Psychologial Research*. url: https :// replicationindex . com / 2016 / 06 / 30 / wagenmakers - default - prior - is - inconsistent - with - the - observed - results-in-psychologial-research (visited on 29. 05. 2019).

Schimmack, Ulrich (2. März 2018a). *Conduct Your Own Replicability Analysis*. Spreadsheet: https://replicationindex.files.wordpress.com/2018/03/spreadsheet-coding1.xlsx. url: https://replicationindex.com/2018/03/02/conduct- your-own-replicability-analysis (visited on 29. 05. 2019).

Schimmack, Ulrich (3. Mai 2018b). *Confused about Effect Sizes? Read more Cohen (and less Loken and Gelman)*. url: https : / / replicationindex . com / 2018 / 05 / 03 / confused - about-statistics-read-more-cohen-and-less-loken-gelman.

Schimmack, Ulrich (4. Dez. 2018c). *Dr. Schnall's R-Index*. url: https://replicationindex. com/tag/schnall.

Schimmack, Ulrich (24. Apr. 2018d). *Estimating Reproducibility of Psychology (No. 151): An Open Post-Publication Peer-Review*. url: https://replicationindex.com/2018/04/24/ estimating - reproducibility - of - psychology - no - 151 - an - open - post - publication-peer-review (visited on 25. 05. 2019).

Schimmack, Ulrich (25. Mai 2018e). *The Fallacy of Placing Confidence in Bayesian Salvation*. url: https : / / replicationindex . com / category / bayes - factor (visited on 29. 05. 2019).

Schimmack, Ulrich & Jerry Brunner (4. Mai 2017a). *How replicable are statistically significant results in social psychology? A replication and extension of Motyl et al. (in press)*. url: https: //replicationindex.com/2017/05- /04/how-replicable-are-statistically- significant - results - in - social - psychology - a - replication - and - extension-of-motyl-et-al-in-press (visited on 29. 05. 2019).

Schimmack, Ulrich & Jerry Brunner (Nov. 2017b). *Z-Curve: A Method for the Estimating Replicability Based on Test Statistics in Original Studies*. url: https://replicationindex. files.wordpress.com/2017/11/adv-meth-practices- -draft-v17-12-08.pdf (visited on 29. 05. 2019).

Schindler, David et al. (2018-06-15). *randomizeR: Randomization for Clinical Trials [version 1.4.2]* [Vignettes 'Assessment and Implementation of Randomization in Clinical Trials', 'Comparing randomizaton procedures', 'Desirability Index', 'randomizeR Quick Reference Guide']. Techn. Ber. CRAN. url: https :// cran .r- project . org / web / packages / randomizeR (visited on 27. 06. 2019).

Schmidt, Heinrich (1991). *Philosophisches Wörterbuch (Kröners Taschenausgabe)* (German Edition). Neu bearbeitet von Georgi Schischkoff. 22. Auflage. Stuttgart: A. Kröner Verlag. isbn: 3520013223.

Schnäbele, Paul (2011). *Bayes-Statistik und konjugierte Verteilungen* [Bachelorarbeit, 20I!)-05-I!)]. Techn. Ber. http://ekp-invenio.physik.uni- karlsruhe.de/record/48253/ files/iekp-bachelor-ka2011-17.pdf. Karlsruhe: Karlsruher Institut für Technologie (KIT). Fakultät für Physik. Institut für Experimentelle Kernphysik.

Schneider, Antonius, Geert-Jan Dinant & Joachim Szecsenyi (2006). Zur Notwendigkeit einer abgestuften Diagnostik in der Allgemeinmedizin als Konsequenz des Bayes'schen Theorems. *Zeitschrift für ärztliche Fortbildung und Qualität im Gesundheitswesen, 100*, 121–127. url: https://www.am.med.tum.de/sites/www.am.med.tum.de-/files/Stufendiagnostik_Bayes_0.pdf (visited on 02. 07. 2020).

Schneider, Carsten Q. & Claudius Wagemann (2012). *Set-Theoretic Methods for the Social Sciences: A Guide to Qualitative Comparative Analysis (QCA).* Heidelberg: Cambridge University Press.

Scholz, Gerhard (1992). *Vipassana Meditation und Drogensucht*. Eine Studie über den Ausstieg aus der Herrschaft der Attraktion Droge. Unveröffentlichte Dissertation, Universität Zürich.

Schönbrod, Felix D. & Eric-Jan Wagenmakers (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin Review, 25*, 128–142. https://link.springer.com/ content/pdf/10.3758/s13423-017-1230-y.pdf.

Schönbrodt, Felix D. (23. Aug. 2013). *Exploring the robustness of Bayes Factors: A convenient plotting function*. url: https :// www . nicebread . de / exploring - the - robustness - of - bayes-factors-a-convenient-plotting-function-2/ (visited on 29. 05. 2019).

Schuirmann, Donald J. (1987). A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics, 15*(6), 657–680.

Schumann, Hans W. (1999). *Der historische Buddha. Leben und Lehre des Gotama*. München: Diederichs.

Schützenmeister, A., U. Jensen & H.-P. Piepho (2012). Checking Normality and Homoscedasticity in the General Linear Model Using Diagnostic Plots. *Communications in Statistics - Simulation and Computation 41*(2), 141–154.

Schwarz, Gideon E. (1978). Estimating the dimension of a model. Annals of Statistics 6(2), 461–464. https://projecteuclid .org/download/pdf_1/euclid.aos/1176344136

Seawright, Jason (2005). Qualitative Comparative Analysis vis-à-vis Regression. *Studies in Comparative International Development, 40*(1), 3–26. https://www.uzh.ch/cmsssl/suz/dam/jcr:00000000-5103-bee3-0000-000022716bb7/05.26.seawright.pdf

Sedlmeier, Peter & Gerd Gigerenzer (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 107*, 309–316.

Sellke, Thomas, M.J. Bayarri & James O. Berger (2001). Calibration of p values for Testing precise Null Hypotheses. *The American Statistician 55*(1), 62–71. url: http : / / www . dcscience.net/Sellke-Bayarri-Berger-calibration-of-P-2001.pdf.

Senn, S. (2001). OPINION — Two cheers for P-values? *Journal of Epidemiology and Biostatistics 6*(2), 193–204. url: https://www.stat.washington.edu/peter/342/Senn.pdf (visited on 04. 05. 2019).

Shafer, Glenn (1982). Lindley's Paradox. *Journal of the American Statistical Association, 77*(378), 325–334.

Shakespeare, William (2014). *Macbeth* [Zweisprachige Ausgabe. I0. Auflage. Übersetzt von Frank Günther. Englische Ausgabe I!)84]. München: dtv.

Shannon, Claude Elwood (1948). A mathematical theory of communication. *The Bell System Technical Journal 27*, 379–425, 623–656.

Shelly, A.L. & E.E. Sibert (1992). Qualitative Analyse: Ein computerunterstützter zyklischer Prozeß. In: Günter L. Huber (Ed.), *Qualitative Analyse. Computereinsatz in der Sozialforschung (*pp. 71–114*).* München: Oldenbourg.

Siebert, Michael & David Ellenberger (2019). Validation of automatic passenger counting: introducing the t-test-induced equivalence test. *Transportation*, 1–15. url: https : / / link.springer.com/article/10.1007%5C%2Fs11116-019-09991-9 (visited on 29. 06. 2019).

Silverfish https://stats.stackexchange.com/users/22228/silverfish, user (9. März 2016). *How much do we know about p-hacking ïn the wild? Cross Validated.* url: https://stats.stackexchange.com/questions/200745/how- much- do- we-know-about-p-hacking-in-the-wild.

Simmons, Joseph P., Leif D. Nelson & Uri Simonsohn (Okt. 2011). False-Positive Psychology. *Psychological Science 22*(11), 1359–1366. ISSN: 1467-9280. doi: 10.1177/0956797611417632. url: http://dx.doi.org/10.1177-/0956797611417632.

Simmons, Joseph P. & Uri Simonsohn (2017). Power Posing: p-Curving the Evidence. *Psychological Science 28*(5), 687–693. Pre-print: http://datacolada.org/wp- content/uploads/ 2017/11/Power-Posing-p-curve-published.pdf.

Simon, Julian Lincoln, Hrsg. (1997). *Resampling: The New Statistics* (2nd ed.) http://www. resample.com/intro-text-online. Arlington/ Virginia: Resampling Stats.

Simons, Daniel J., Alex O. Holcombe & Barbara A. Spellman (2014). An Introduction to Registered Replication Reports at Perspectives on Psychological Science. *Perspectives on Psychological Science 9*(5), 552–555. https://journals. sagepub.com/doi/pdf/10 .1177 /1745691614543974.

Simonsohn, Uri (2014). *Posterior-Hacking: Selective Reporting Invalidates Bayesian Results Also* (published 20I4-02-0I). Online Supplementary Materials: http://urisohn.com/sohn_files/ papers / Supp _ Post _ Hacking . pdf. url: https : / / papers . ssrn . com / sol3 / papers.cfm?abstract_id = 2374040.

Simonsohn, Uri, Leif D. Nelson & Joseph P. Simmons (2013). *P-curve Online app | Official User- Guide to the P-curve*. Online app: http://www.p- curve.com/app4, user guide: http://www.p-curve.com/guide.pdf. (Visited on 22. 06. 2019).

Simonsohn, Uri, Leif D. Nelson & Joseph P. Simmons (2014). P-curve: A key to the File Drawer. *Journal of Experimental Psychology: General 143*(2), 534–547. issn: 0096-3445. doi: 10.1037/a0033242. url: http://dx.doi.org/10.1037-/a0033242.

Simonsohn, Uri, Leif D. Nelson & Joseph P. Simmons (2014-01-11). *P-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results* [last revised 2014-11-19. available at SSRN. url: https :// dx . doi . org / 10 . 2139 / ssrn . 2377290 (visited on 22. 06. 2019).

Simpson, Edward H. (1951). The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society. Series B 13*(2), 238–241. https://math.bme.hu/ ~ marib/bsmeur/ simpson.pdf.

Skinner, Burrhus Frederic (1948). 'Superstition' in the Pigeon. *Journal of Experimental Psychology 38*, 168–172. url: https://psychclassics.yorku.ca/Skinner/Pigeon.

Smith, John K. & Louis Heshusius (1986). Closing Down the Conversation: The End of the Quantitative-Qualitative Debate Among Educational Inquirers. *Educational Researcher 15*(1), 4–12.

Snakes, Echidne of the (26. Okt. 2012). *Women Vote Their Hormones: The Study itself*. url: http://echidneofthesnakes .blogspot.com/2012/10/women-vote-their-hormones-study-itself.html (visited on 26. 10. 2004).

Snodgrass, Steve (2011-09-27). *Examining Retroactive Facilitation of Recall: an Adapted Replication of Bem (2011, Study 9) and Galak and Nelson (2010)*. SSRN. https://papers. ssrn.com/sol3/papers.cfm?abstract_id = 1935942.

Solé-Leris, Amadeo (1994). *Die Meditation, die der Buddha selber lehrte*. Freiburg im Breisgau: Spektrum.

Spanos, Aris (2013). Who Should Be Afraid of the Jeffreys-Lindley Paradox? *Philosophy of Science, 80*(1), 73–93. doi: 10.1086/668875.

Sparrer, Insa & Matthias Varga v. Kibéd (2000). *Ganz im Gegenteil. Für Querdenker und solche, die es werden wollen. Tetralemmaarbeit und andere Grundformen Systemischer Strukturaufstellungen*. Heidelberg: Carl-Auer Systeme.

Spiegelhalter, David J. (2004). Incorporating Bayesian ideas into health-care evaluation. *Statistical Science 19*(1), 156–174. https://projecteuclid.org/euclid.ss/1089808280.

Spiegelhalter, David J., Nicola G. Best et al. (2002). Bayesian Measure of Model Complexity and Fit. *Journal of the Royal Statistical Society. Series B. Statistical Methodology 64*(4), 583–639. https://rss.onlinelibrary.wiley.com/doi/epdf/10 .1111 /1467-9868.00353.

Spiegelhalter, David J., Nicola G. Best et al. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society. Series B. Statistical Methodology 76*(3), 485– 493.

Spörlein, Christoph (2021). *Text Mining of Social Media data using R*. url: https : / / www .cspoerlein.com/files/textanalyse.html (visited on 24. 11. 2021).

Sprenger, Jan (2012). Testing a Precise Null Hypothesis: The Case of Lindley's Paradox. *Philosophy of Science, 80*, 733–744. http://philsci-archive.pitt.edu/9419/1/LindleyPSA.pdf.

Sprenger, Jan (2018). The Objectivity of Subjective Bayesian Inference. *European Journal for Philosophy of Science 8*, 539–558. url: http://philsci-archive.pitt.edu/11797/ 1/ObjectiveBayesianStatistics.pdf (visited on 23. 05. 2019).

Stachowske, Ruthard (2002). *Mehrgenerationentherapie und Genogramme in der Drogenhilfe. Drogenabhängigkeit und Familiengeschichte*. Heidelberg: Asanger.

Stanger, Charles, Rajiv Jhangiani & Hammond Tarry (2014). *Principles of Social Psychology-1st International Edition*. https : / / opentextbc . ca / socialpsychology, adapted from Charles Stagnor (2011), Principles of Social Psychology. British Columbia: B.C. Open Textbook Project, BCampus Open Education.

Stangl, Werner (o. D.). *Vergleich der Schlußformen Abduktion, Induktion und Deduktion* [Arbeitsblätter]. url: https://arbeitsblaetter.stangl-taller.at/DENKENTWICKLUNG/ Abduktion-Induktion-Deduktion.shtml.

Stansbury, Dustin (21. Apr. 2013). *Model Selection: Underfitting, Overfitting, and the Bias-Variance Tradeoff. The Clever Machine. Topics in Computational Neuroscience and Machine Learning*. url: https :// theclevermachine . wordpress . com / 2013 / 04 / 21 / model - selection-underfitting-overfitting-and-the-bias-variance-tradeoff/ (visited on 29. 05. 2019).

Statista (2019). *Verschiedene Statistiken rund um Tee*. Statista GmbH. url: https : / / de . statista.com (visited on 22. 05. 2019).

Stefan, Angelika M. u. a. (2019). A Tutorial on Bayes Factor Design Analysis with Informed Priors [last version: 2018-07-02. *Behavior Research Methods 51*(3). Pre-print from 2019-06-04: https : / / doi . org / 10 . 31234 / osf . io / aqr79, supplemental material: https : / / osf . io / n8m4e, S. 1042–1058. url: https : / / link . springer . com / content / pdf/10.3758%5C%2Fs13428-018-01189-8.pdf.

Stefan Angelika & Schönbrodt, Felix D. (10. Jan. 2017). *Two meanings of priors, part I: The plausibility of models*. url: https://www.nicebread.de/two-meanings-of-priors-1 (visited on 05. 06. 2019).

Stegmüller, Wolfgang (1973a). *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie. Band IV. Personelle und Statistische Wahrscheinlichkeit. Studienausgabe Teil A. Aufgaben und Ziele der Wissenschaftstheorie. Induktion. Das ABC der odernen Wahrscheinlichkeitstheorie und Statistik*. Berlin: Springer.

Stegmüller, Wolfgang (1973b). *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie. Band IV. Personelle und Statistische Wahrscheinlichkeit. Studienausgabe Teil C. Carnap II: Normative Theorie des induktiven Räsonierens*. Berlin: Springer.

Stegmüller, Wolfgang (1975). *Das Problem der Induktion: Humes Herausforderung und moderne Antworten. Der sogenannte Zirkel des Verstehens*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Stein, Rosemarie (1999-04-23). Früherkennungsprogramme/Epidemiologie: Testergebnisse richtig interpretieren. *Deutsches Ärzteblatt, 96*(16), A-1044 / B-869 / C–813. url: https:// www.aerzteblatt.de/pdf.asp?id = 16819 (visited on 22. 05. 2019).

Steinke, Ines (2000). Gütekriterien qualitativer Forschung. In Uwe Flick, Ernst v. Kardorff & Ines Steinke (Eds.), *Qualitative Forschung. Ein Handbuch* (pp. 319–331). Reinbek bei Hamburg: Rowohlts Enzyklopädie..

Sterling, T.D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance: Or vice versa. *Journal of the American Statistical Association 54*(285), 30–34.

Sterling, T.D., W.L. Rosenbaum & J.J. Weinkam (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice-versa. *American Statistician 49*(1), 108–112. url: https://www.gwern.net/docs/statistics/bias/1995-sterling.pdf (visited on 24. 06. 2019).

Stigler, Stephen M. (2007). The Epic Story of Maximum Likelihood. *Statistical Science 22*(4),598–620. https://arxiv.org/pdf/0804.2996.pdf.

Stigler, Stephen M. (2013). The True Title of Bayes' Essay. *Statistical Science 28*(3), 283–288. https://arxiv.org/pdf/1310.0173.pdf.

Stoner, J.A. (1961). *A comparison of individual and group decision involving risk*. Unpublished Master's Thesis. Cambridge/ Massachusetts: Massachusetts Institute of Technology.

Strauss, Anselm & J. Corbin (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, CA: Sage.

Stucchio, Chris (2014a). *Asymptotics of Evan Miller's Bayesian A/B formula*. url: https : / /www . chrisstucchio . com / blog / 2014 / bayesian _ asymptotics . html (visited on 09. 06. 2014).

Stucchio, Chris (2014b). *Easy Evaluation of Decision Rules in Bayesian A/B testing*. url: https:// www . chrisstucchio . com / blog / 2014 / bayesian _ ab _ decision _ rule . html (visited on 05. 06. 2014).

Stucchio, Chris (2015). *Bayesian A/B Testing at VWO*. Technical white paper. url: https :// cdn2.hubspot.net/hubfs/310840/VWO_SmartStats_technical_whitepaper. pdf (visited on 02. 09. 2015).

Studer, Urban M. (1995). *Therapieforschung in STARTAGAIN. Eine Werkstattschau*. Jahresbericht des START AGAIN 1995. Zürich.

Studer, Urban M. (1996a). *Evaluation des Suchttherapiezentrums Start Again*. Zwischenbericht an das Bundesamt für Justiz (BAJ) vom Dezember 1996. Zürich.

Studer, Urban M. (1996b). *Wahrscheinlichkeit als Logik: Die formale Struktur konsistenten Schlussfolgerns*. Zwischenbericht an das Bundesamt für Justiz (BAJ) vom Dezember 1996. Zürich.

Studer, Urban M. (1998). *Verlangen, Süchtigkeit und Tiefensystemik*. Techn. Ber. Evaluationsbericht an das Justizministerium (BAJ) der Schweiz, http://www.ofj.admin.ch/themen/ stgb-smv/ber-mv/37.pdf. Zürich.

Studer, Urban M. (2005). *Berufliche Massnahmen in der IV bei suchtkranken Menschen — Follow- up*. Vortrag bei SVA Zürich am 23.0!).2005. Zürich.

Su, Yu-Sung et al. (2011). Multiple Impuation with Diagnostics (mi) in R: Opening Windows into the Black Box. *Journal of Statistical Software 45*(2), 1–3. http :// www . stat. columbia . edu/ ~ gelman/research/published/mipaper.pdf1.

Susanti, Yuliana et al. (2014). M estimation, S estimation, and MM estimation in robust regression. *International Journal of Pure and Applied Mathematics 91*(3), 349–360. https : / / www . researchgate . net / publication / 266599562 _M_ estimation _S_ estimation _ and_MM_estimation_in_robust_regression. (Visited on 14. 10. 2020).

Sverdlov, Oleksandr, Yevgen Ryeznik & Sheng Wu (2015). Exact Bayesian Inference Comparing Binomial Proportions, With Application to Proof-of-Concept Clinical Trials. *Therapeutic Innovation and Regulatory Science 490*(1). R-Code: https://journals.sagepub.com / doi / suppl / 10 . 1177 / 2168479014547420 / suppl _ file / DS _ 10 . 1177 _ 2168479014547420.zip. (Visited on 10. 07. 2019).

Szucs, D. & J.P. Ioannidis (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS biology 15*(3), 163–174. https://www.ncbi.nlm.nih.gov/pmc-/articles/PMC5333800/.

Szucs, D. & J.P. Ioannidis (2017-08-03). When Null Hypothesis Significance Testing is unsuitable for research: A reassessment. *Frontiers in Human Neuroscience 11*. 21 pages, https://www.frontiersin.org/articles/10.3389-/fnhum.2017.00390/full.

Taboga, Marco (2010). *Bayesian estimation of the parameters of the normal distribution* [Lectures on probability and statistics]. url: https : / / www . statlect . com / fundamentals - of - statistics / normal - distribution - Bayesian - estimation (visited on 24. 05. 2019).

Tajfel, H. (1970). Experiments in intergroup discrimination. *Scientific American 223*, 96– 102.

Taleb, Nassim Nicholas (2007). *The Black Swan. The Impact of the Highly Improbable*. New York: Random House.

Tashakkori, Abbas & Charles Teddlie, Hrsg. (1998). *Applied social research methods series, Vol. 46. Mixed methodology: Combining qualitative and quantitative approaches*. Thousand Oaks: Sage.

Tashakkori, Abbas & Charles Teddlie, Hrsg. (2003a). *Handbook of mixed methods in social and behavioral research*. Thousand Oaks: Sage.

Tashakkori, Abbas & Charles Teddlie (2003b). The past and future of mixed methods research from data triangulation to mixed model designs. In: Abbas Tashakkori & Charles Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 671–702). Thousand Oaks: Sage.

Tate, Charlotte Ursula (Juni 2015). *Type III and Type IV Errors: Statistical Decision-Making Considerations in addition to Rejecting and Retaining the Null Hypothesis*. url: http :// web . stanford.edu/group/bps/cgi- bin/wordpress/wp-content/uploads/2015/ 06/Tate.pdf.

Team, Analytics Vidhya Content (4. März 2016). *Tutorial on 5 Powerful R Packages used for imputing missing values*. Analytics Vidhya — Learn everything about analytics. url: https://www.analyticsvidhya.com/blog/2016/03-/tutorial-powerful-packages- imputing-missing-values (visited on 23. 06. 2019).

Team, The JAGS Development (2019). *JAGS — Just Another Gibbs Sampler*. url: http://mcmc- jags.sourceforge.net (visited on 21. 05. 2019).

Team, The OPENBUGS Development (2019). *OpenBUGS — Bayesian inference Using Gibbs Sampling [open-source version of BUGS]*. url: http://www.openbugs.net (visited on 21. 05. 2019).

Teddlie, Ch. und A. Tashakkori (2006). A general typology of research designs featuring mixed methods. *Research in the Schools 13*(1), 12–28.

Teesalon, Berliner (2017). *Teeanbaugebiete in Indien*. url: http :// www . tee - import . de / indien/karte.htm (visited on 22. 05. 2019).

Teeverband, Deutscher (2017). *TEE als Wirtschaftsfaktor*. Techn. Ber. Hamburg. url: https://www.teeverband.de-/fileadmin/user_upload/WFT_2017_DE.pdf (visited on 22. 05. 2019).

Teeverband, Deutscher (2018). *TEE als Wirtschaftsfaktor*. Techn. Ber. Hamburg. url: https://www.teeverband.de/fileadmin-/Redaktion-/WFT/WFT_2018/WFT_fin3_klein.pdf.

The Jasp Development Team [Wagenmakers, Eric-Jan u. a. (2018). *JASP* [latest version: 20I8-I2-I0, v0.1.2.0]. url: https://jasp-stats.org/team (visited on 06. 05. 2019).

The men who stare at goats (2009).

The Stan Development Team: Gelman, Andrew et al. (2019a). *Documentation. Stan User's Guide. Stan Language Reference Manual. Stan Language Functions Reference. RStan Documentation. PyStan Documentation. Case Studies and Notebooks. Tutorials. Specialized Field Guides. The Stan Forums. GitHub Stan Developer Wiki. Further References*. Stan Governing Body | NumFOCUS. url: https : / / mc - stan . org / users / documentation (visited on 05. 06. 2019).

The Stan Development Team: Gelman, Andrew et al. (2019b). *Stan - state-of-the-art platform for statistical modeling and high-performance statistical computation*. url: https://mc-stan. org (visited on 21. 05. 2019).

Thiem, Alrik (2012). Unifying Configurational Comparative Methodology. Generalized-Set Qualitative Comparative Analysis. *Political Methodology Committee on Concepts and Methods. Working Paper Series 34*. https : / / www . concepts - methods . org / Files /WorkingPaper/PM_34_Thiem.pdf.

Thiem, Alrik & Michael Baumgartner (30. Juni 2016). *Glossary for Configurational Comparative Methods [Version I.I]*. url: https : / / journal . r - project . org / archive / 2013 - 1/thiem-dusa.pdf (visited on 20. 03. 2019).

Thiem, Alrik & Adrian Dus, a (2013a). QCA: A Package for Qualitative Comparative Analysis. *The R Journal 5*(1). https://journal.r-project.org/archive/2013-1/thiem- dusa.pdf.

Thiem, Alrik & Adrian Dus, a (2013b). *Qualitative Comparative Analysis with R: A User's Guide*. New York: Springer.

Thomas, Samuel & Wanzhu Tu (2020). *Learning Hamiltonian Monte Carlo in R [latest version: 2020-I2-2I]*. https :// arxiv . org / pdf / 2006 . 16194 . pdf. Indiana/ USA: Department of Biostatistics and Health Data Science, Indiana University School of Medicine. (Visited on 21. 12. 2020).

Thulin, Måns (2014). *On Confidence Intervals and Two-Sided Hypothesis Testing*. Diss. Uppsala/ Schweden: Department of Mathematics, Uppsala University.

Thulin, Måns (2015-05-28). *Is rejecting the hypothesis using p-value equivalent to hypothesis not belonging to the confidence interval?* [Answer to a blogpost]. https : / / stats . stackexchange . com / questions / 169141 / is - rejecting - the - hypothesis - using-p-value-equivalent-to-hypothesis-not-belonging. Diss.

Thulin, Måns & Silvelyn Zwanzig (2017). Exact confidence intervals and hypothesis tests for parameters of discrete distributions. *Bernoulli 23*(1), 479–502. https://arxiv.org/pdf/1412. 0442.

Tim https://stats.stackexchange.com/35989/tim, user (10. März 2016). *Do Bayesian priors become irrelevant with large sample size? Cross Validated*. url: https://stats. stackexchange . com / questions / 200982 / do - bayesian - priors - become - irrelevant-with-large-sample-size (visited on 30. 07. 2020).

Tinto, V. (1986). Theories of college student departure revisited. In John C. Smart (Ed.), *Higher education: Handbook of theory and research Vol.2 (pp.* 359–384). New York: Agathon.

Titanica, Encyclopedia (1996). *RMS Titanic facts, history and passenger and crew biography*. url: https://www.encyclopedia-titanica.org/titanic (visited on 20. 05. 2019).

Tofallis, Chris (2008). Least Squares Percentage Regression. *Journal of Modern and Applied Statistical Methods, 7*. https://papers.ssrn.com/sol3/papers.cfm?abstract_ id = 1406472, S. 526–534. (Visited on 14. 10. 2020).

Trafimow, David, Valentin Amrhein u. a. (2018). Manipulating the Alpha Level Cannot Cure Significance Testing. *Frontiers in Psychology, 9*, 699. https : / / www . frontiersin . org / article/10.3389/fpsyg.2018.00699.

Trafimow, David & Michael Marks (2015). Editorial. *Basic and Applied Social Psychology 37*(1), 1–2. https :// www . tandfonline . com / doi / full / 10 . 1080 / 01973533 . 2015 . 1012991.

Traxler, Matthew J. u. a. (2012). Feeling the past: The absence of experimental evidence for anomalous retroactive influences on text processing. *Memory and Cognition 40*(8), 1366–1372. https://link.springer.com/content/pdf-/10.3758%2Fs13421- 012- 0232- 2.pdf.

Tripepi, Giovanni et al. (2010). Selection bias and information bias in clinical research. *Nephron Clin. Pract., 115*, c94–c99. Published online 2010-04-21, https://www.karger.com/Article/ Pdf/312871.

Tristl, Christiane, Martin Müller & Veit Bachmann (2015). Lexicometric Analysis: A Methodological Prelude. In: Veit Bachmann & Martin Müller (Eds.), *Perceptions of the EU in Eastern Europe and Sub-Saharan Africa. Europe in a Global Context*. London: Palgrave Macmillan.

Tschirk, Wolfgang (2014). *Statistik: Klassisch oder Bayes*. Berlin: Springer-Verlag GmbH. ISBN: 3642543847.

Tu, Chengcheng & Emma K.T. Benn (2017). RRApp, a robust randomization app, for clinical and translational research. *Journal of Clinical and Translational Science 1*(6), 323–327. url: https://www.ncbi.nlm.nih.gov/pmc/articles-/PMC5915802 (visited on 27. 06. 2019).

Tukey, John Wilder (1949). Moments of random group size distributions. *Annals of Mathematical Statistics 20*(4), 523–539.

Tukey, John Wilder (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics [Abstracts of Papers] 29*, 614. https :// projecteuclid . org / download / pdf_1/euclid.aoms/1177706647.

Tukey, John Wilder (1962). The future of data analysis. Annals of Mathematical Statistics, 33(1), 1-67. https://projecteuclid.org/download/pdf_1/euclid.aoms/1177704711.

Tukey, John Wilder (1977). *Exploratory Data Analysis*. Reading/ Massachusetts: Addison-Wesley.

Tukey, John Wilder (1980). We Need Both Exploratory and Confirmatory. *The American Statistician, 34*(1), 23–25.

Tukey, John Wilder (1991). The Philosophy of Multiple Comparisons. *Statistical Science, 6*(1), 100–116.

Turk, James (12. Jan. 2017). *How much did it cost to build the Titanic*. url: https://www.fgmr. com/how-much-did-it-cost–to-build-the-titanic.

U Ko Lay (1995). *Essence of Tipi.taka*. Igatpuri, India: V.R.I.

Uexküll, Jakob Johann von (1920). *Theoretische Biologie*. Berlin: von Gebrüder Paetel.

Ulrich, Rolf et al. (2016). Inflation von falsch-positiven Befunden in der psychologischen Forschung. Mögliche Ursachen und Gegenmaßnahmen. *Psychologische Rundschau, 67*(3), 163–174.

Union, Inter-Parliamentary (2009a). *Women in National Parliaments: World Average [Situation as of 3I March 2009)*]. url: http : / / www . ipu . org / wmn - e / world . htm (visited on 08. 05. 2009).

Union, Inter-Parliamentary (2009b). *Women in National Parliaments: World Classification [Situation as of 3I March 2009)]*. url: http://www.ipu.org/wmn-e/classif.htm (visited on 08. 05. 2009).

unknown (1998). *Titanic Inquiry Project. Electronic copies of the inquiries into the disaster*. Webpage about the Titanic inquiry [including reports]. Titanic Inquiry Project [TIP]. url: https://www.titanicinquiry.org.

unknown (2019). *Presidential Heights and Weights*. url: http : / / www . american - presidents . info / presidential - heights - and - weights . html (visited on 06. 06. 2019).

Uschner, Diane (2018-15-06). *Tutorial: comparing randomization procedures with randomizeR [Vignette]*. Techn. Ber. https : / / cran . r - project . org / web / packages / randomizeR/vignettes/comparison-example.pdf. CRAN.

Uschner, Diane et al. (2018). randomizeR: An R Package for the Assessment and Implementation of Randomization in Clinical Trials. *Journal of Statistical Software 85*(8), 1–22. https://www.ideal.rwth-aachen.de/wp- content/uploads-/2017/11/article_accepted.pdf.

Ustorf, Anne-Ev (2008). *Wir Kinder der Kriegskinder: Die Generation im Schatten des Zweiten Weltkriegs*. Freiburg im Breisgau: Herder.

van Horn, Kevin S. (2004). *Unofficial Errata and Commentary for E. T. Jaynes's Probability Theory: The Logic of Science* [last update: 2004-II-06]. url: http://www.ksvanhorn.com/bayes/ jaynes/jaynes.html.

Vanderplas, Susan & Heike Hofman (2015-08). *Spatial Reasoning and Data Displays* [latest version 2015-10-25]. IEEE Transactions on Visualization and Computer Graphics [online]. doi: 10.1109/TVCG.2015.2469125.

various (27. Juli 2017). What were the ticket prices to board the Titanic? *Quora — wo man Wissen miteinander austauscht und die Welt besser verstehen kann*. url: https :// www . quora . com / What - were - the - ticket - prices - to - board - the - Titanic (visited on 20. 05. 2019).

Vasishth, Shravan et al. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language, 103*, 151–175. http://www.stat.columbia. edu/ ~ gelman/research/published-/VMJG2018.pdf.

Vehtari, Aki, Andrew Gelman & Jonah Gabry (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*(5), 1413–1432. Pre-print [v5, from 2016-09-12]: https://arxiv.org/abs/1507.04544.

Villa, Cristiano & Stephen Walker (2015-03-13). On the mathematics of the Jeffreys-Lindley Para- dox. Techn. Ber. Published 2009: https://arxiv.org/abs/1503.04098v1.

Vipassana Research Institute, Ed. (1993). *Mahasatipaatthana Suttam*. Igatpuri, India: V.R.I.

von Lucadou, Walter (2012). *Die Geister, die mich riefen: Deutschlands bekanntester Spukforscher erzählt* [mit Peter Wagner]. Köln: Bastei Lübbe.

von Ludwiger, Illobrand (2013). *Das neue Weltbild des Physikers Burkhard Heim: Unsterblich in der 6-dimensionalen Welt*. München: Komplett-Media, Grünwald.

Vorderer, Peter & Norbert Groeben (Eds.) (1987). *Textanalyse als Kognitionskritik? Möglichkeiten und Grenzen ideologiekritischer Inhaltsanalyse*. Tübingen: Gunter Narr.

Wagenmakers, Eric–Jan (2007a). A Practical Solution to the Pervasive Problems of p–Values. *Psychonomic Bulletin and Review, 14*(5), 779–804. Corrections: http://www.ejwagenmakers.com/2007/CorrigendumPvalues.pdf. url: https://ejwagenmakers.com/2007/pValueProblems.pdf (visited on 29. 05. 2019).

Wagenmakers, Eric–Jan (2007b). Stopping Rules and Their Irrelevance for Bayesian Inference: Online Appendix to A Practical Solution to the Pervasive Problems of p–Values, to appear in *Psychonomic Bulletin and Review*. url: http://www.ejwagenmakers.com/2007/WretchedPriorAppendix.pdf (visited on 29. 05. 2019).

Wagenmakers, Eric–Jan (2007c). That Wretched Prior: Online Appendix to A Practical Soluti- on to the Pervasive Problems of p–Values, to appear in *Psychonomic Bulletin and Review*. url: http://www.ejwagenmakers.com-/2007/WretchedPriorAppendix.pdf (visited on 29. 05. 2019).

Wagenmakers, Eric–Jan, Michael Lee et al. (2019). *Another Statistical Paradox*. sub- mitted for publication. url: https:/ /ejwagenmakers.com/submitted/AnotherStatisticalParadox.pdf.

Wagenmakers, Eric–Jan, Ruud Wetzels, Denny Borsboom, Rogier Kievit et al. (2011). *Yes, Psychologists Must Change the Way They Analyze Their Data: Clarifications for Bem, Utts, and Johnson (2011). Response to a previous draft of Bem et al.*: https://ejwagenmakers.com/2011/ClarificationsForBemUttsJohnson.pdf.

Wagenmakers, Eric–Jan, Ruud Wetzels, Denny Borsboom & Han L.J. van der Maas (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100*(3), S. 426–432. Online appendix: https://ejwagenmakers.com/2011/OnlineAppendixBem1.pdf. url: http:// dx.doi.org/10.1037/a0022790 (visited on 29. 05. 2019).

Wahl, Diethelm, Hrsg. (2006). *Lernumgebungen erfolgreich gestalten: Vom trägen Wissen zum kompetenten Handeln*. Weinheim: Julius Klinkhardt.

Walker, Esteban & Amy S. Nowacki (2011). Understanding Equivalence and Noninferiority Testing. *Journal of General Internal Medicine, 26*(2), 192–196. url: https://www.ncbi. nlm.nih.gov/pmc/articles/PMC3019319 (visited on 29. 06. 2019).

Wallace, David Foster (1998). *A Supposedly Fun Thing I'll Never Do Again: Essays and Arguments*. New York: Little Brown & Company Inc.

Waller, Niels G. (2004). The fallacy of the null hypothesis in soft psychology. *Applied and Preventive Psychology 11*, 83–86.

Walters, Elizabeth J., Christopher H. Morrell & Richard E. Auer (2006). An Investigation of the Median-Median Method of Linear Regression. *Journal of Statistics Education, 14*(2). http://jse.amstat.org/v14n2/morrell.html.

Wang, Min & Guangying Liu (2016). A simple two-sample Bayesian t-Test for hypothesis testing. *The American Statistician 70*(2), 195–201. url: https://www.researchgate.net/publication-/281670776_A_Simple_Two-Sample_Bayesian_t-Test_for_ Hypothesis_Testing (visited on 24. 07. 2019).

Wasserstein, Ronald L. (26. Feb. 2015). *ASA comment on a journal's ban on nullhypothesis statistical testing.* url: https : / / community . amstat . org / blogs / ronald - wasserstein / 2015 / 02 / 26 / asa - comment - on - a - journals - ban - on - null - hypothesis - statistical-testing.

Wasserstein, Ronald L. & Nicole A. Lazar (2016). The ASA's Statement onp-Values: Context, Process, and Purpose. *The American Statistician, 70*(2), 129–133. http :// dx .doi . org / 10 . 1080/00031305.2016.1154108.

Watanabe, Sumio (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research, 11*, 3571–3594. http://jmlr.csail.mit.edu/papers /volume11/watanabe10a/watanabe10a.pdf.

Watanabe, Sumio (2013). A Widely Applicable Bayesian Information Criterion. *Journal of Machine Learning Research 14*, 867–897. url: http : / / www . jmlr . org / papers / volume14/watanabe13a/watanabe13a.pdf (visited on 17. 07. 2020).

Watzlawick, Paul, Janet H. Beavin & Don D. Jackson (1974). *Menschliche Kommunikation. Formen, Störungen, Paradoxien*. Bern: Verlag Hans Huber.

Weber, Max (1904/1991). *Die „Objektivität sozialwissenschaftlicher und sozialpolitischer Erkenn nis. Schriften zur Wissenschaftslehre [1991]*. Erstdruck Archiv für Sozialwissenschaft und Sozialpolitik, 19. Bd., Heft 1, 1904, 22–89. http://www.zeno.org/Soziologie/M/Weber + Max/Schriften + zur + Wissenschaftslehre/Die + %C2%BBObjektivit% C3 % A4t % C2 % AB + sozialwissenschaftlicher + und + sozialpolitischer + Erkenntnis. Stuttgart: Reclam.

Weber, Nina (2012-03-15). *Wahrsage-Studie. Psychologen sehen schwarz fürs Hellsehen*. https:// www.spiegel.de-/wissenschaft/mensch/hellsehen-studie-von-daryl-bem- nicht-reproduziert-a-821355.html. Hamburg: Der Spiegel Online.

Weber, Robert Philip (1990). *Basic content analysis. Sage University Papers Series on Quantitative Applications in the Social Sciences, series no. 07-049*. Quantitative applications in the social sciences Retrieved from WorldCat database Retrieved from http://catdir.loc.gov/catdir/enhancements/fy0655/!)006I0I!)-d. html.

Weinberger, Joel. L. (1994). Can personality change? In: Todd F. Heatherton & Joel L. Weinberger (Eds.), *Can personality change?* (pp. 333–350). Washington, D.C.: American Psychological Association.

Weise, Tobias (2014). *Coincidence Analysis (CNA): Reproducing Baumgartner and Epple 20I4 in R [20I4-08-21]*. url: https : / / tobiasweise . de / 2014 / 08 / 21 / reproducing - baumgartner-and-epple-2014-in-r.html.

Weisstein, Eric W. (o.J.). *Lyapunov Condition*. http : / / mathworld . wolfram . com / LyapunovCondition.html.

Welbers, Kasper, Wouter Van Atteveldt & Kenneth Benoit (2017). Text analysis in R. *Communication Methods and Measures, 11*(4), 245–265.
Welch, B.L. (1947). The generalizations of 'Student's' problem when several different population variances are involved. *Biometrika, 34*(1–2), 28–35. http://pds9.egloos.com/pds/200804/ 26/44/2332510.pdf.
Welch, B.L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika, 38*(3–4), 330–336. https://www.jstor.org/stable/2332579.
Wellek, Stefan (2010). *Testing statistical hypotheses of equivalence and noninferiority*. London: Chapman, HallChapman & Hall/CRC/CRC.
Wernet, Andreas (2000). *Einführung in die Interpretationstechnik der Objektiven Hermeneutik*. Opladen: Leske + Budrich.
West, Stuart A., S. Griffin Ashleigh & Andy Gardner (2007). Evolutionary explanations for cooperation. *Current Biology, 17*(16), R661–R672. https://www.sciencedirect.com/science/ article/pii/S0960982207014996.
Whitehead, Alfred N. & Bertrand Russell (1910). *Principia Mathematica*. https :// quod . lib.umich.edu/cgi/t/text/text-idx?c = umhistmath;idno = AAT3201.0001. 001. Cambridge: Cambridge University Press.
Wikipedia (Ed.) (2019a). *Wikipedia: Die freie Enzyklopädie. AIDS#Deutschland*. St. Petersburg, Florida (USA): Wikimedia Foundation Inc. url: http :// de . wikipedia . org / wiki / AIDS%5C#Deutschland (visited on 23. 05. 2019).
Wikipedia (Ed.) (2019b). *Wikipedia: Die freie Enzyklopädie. Datei:Bayes' Theorem 2D.svg*. St. Petersburg, Florida (USA): Wikimedia Foundation Inc. url: http : / / de . wikipedia . org/wiki/AIDS%5C#Deutschland (visited on 23. 05. 2019).
Wikipedia (Ed.) (2019c). *Wikipedia: freie Enzyklopädie. RMS Titanic*. St. Petersburg, Florida (USA): Wikimedia Foundation Inc. url: https :// de . wikipedia . org / wiki / RMS _ Titanic (visited on 20. 05. 2019).
Wikipedia (Ed.) (2019d). *Wikipedia: freie Enzyklopädie. Weltbevölkerung*. St. Petersburg, Florida (USA): Wikimedia Foundation Inc. url: https : / / de . wikipedia . org / wiki / Weltbev%5C%C3%5C%B6lkerung (visited on 16. 06. 2019).
Wikipedia (Ed.) (2019e). *Wikipedia: The Free Encyclopedia. Conjugate prior*. St. Petersburg, Florida (USA): Wikimedia Foundation Inc. url: https : / / en . wikipedia . org / wiki / Conjugate_prior (visited on 23. 05. 2019).
Wikipedia (Ed.) (2019f). *Wikipedia: The Free Encyclopedia. Crew of the RMS Titanic*. St. Petersburg, Florida (USA): Wikimedia Foundation Inc. url: https :// en . wikipedia . org / wiki/Crew_of_the_RMS_Titanic (visited on 20. 05. 2019).
Wikipedia (Ed.) (2019g). *Wikipedia: The Free Encyclopedia. Heights of presidents and presidential candidates of the United States*. St. Petersburg, Florida (USA): Wikimedia Foundation Inc. url: https://en.wikipedia.org/wiki/ Heights_of_presidents_and_presidential_candidates_of_the_United_States (visited on 06. 06. 2019).
Wikipedia (Ed.) (2019h). *Wikipedia: The Free Encyclopedia. List of average human height world- wide*. St. Petersburg, Florida (USA): Wikimedia Foundation Inc. url: https : / / en . wikipedia .org / wiki / List _ of _ average _ human _ height _ worldwide (visited on 06. 06. 2019).
Wikipedia (Ed.) (2019i). *Wikipedia: The Free Encyclopedia. Maximum entropy probability distribution*. St. Petersburg, Florida (USA): Wikimedia Foundation Inc. url: https : / / en . wikipedia . org / wiki / Maximum _ entropy _ probability _ distribution % 5C # Uniform_and_piecewise_uniform_distributions (visited on 06. 06. 2019).
Wikipedia (Ed.) (2019j). *Wikipedia: The Free Encyclopedia. Metropolis–Hastings algorithm*. St. Petersburg, Florida (USA): Wikimedia Foundation Inc. url: https :// en . wikipedia . org / wiki / Metropolis % 5C % E2 % 5C % 80 % 5C % 93Hastings _ algorithm (visited on 04. 06. 2019).
Wikipedia (Ed.) (2020). *Wikipedia: The Free Encyclopedia. Lindley's paradox*. St. Petersburg, Florida (USA): Wikimedia Foundation Inc. url: https://en.wikipedia.org/wiki/ Lindley%27s_paradox (visited on 14. 10. 2020).
Wilcox, Rand R. (1989). Comparing the variances of dependent groups. *Psychometrika, 54*, 305–315.
Wilcox, Rand R. (1995). *Statistics for the social sciences*. San Diego: Academic Press.
Wilcox, Rand R. (2012). *Introduction to robust estimation and hypothesis testing*. San Diego: Academic Press.
Wilkinson, Darren (2004a). *A simple Gibbs sampler [partof: Implementation of stochastic simulation algorithms, MAS45I: Principles of statistics. Part 2: Stochastic simulation and MCMC]*. Further infos: https://www.mas.ncl.ac.uk-/ ~ ndjw1/teaching/sim/gibbs. url: https://www.mas.ncl.ac.uk/ ~ ndjw1/teaching/sim/gibbs/gibbs.html (visited on 05. 06. 2019).
Wilkinson, Darren (2004b). *Metropolis-Hastings sampling*. Further infos: https://www.mas. ncl . ac . uk / ~ ndjw1 / teaching / sim / metrop / metrop . html, https : / / www . mas . ncl . ac . uk / ~ ndjw1 / teaching / sim / metrop / indep . html. url: https://www.mas.ncl.ac.uk/ ~ ndjw1/teaching/sim/metrop (visited on 05. 06. 2019).
Wilkinson, Darren (18. Aug. 2010). *Metropolis-Hastings MCMC algorithms*. url: https : / / darrenjw . wordpress . com / 2010 / 08 / 15 / metropolis - hastings - mcmc - algorithms (visited on 05. 06. 2019).
Wilkinson, Leland and the Task Force on Statistical Inference. APA Board of Scientific Affairs (1999). Statistical Methods in Psychology Journals. Guidelines and Explanations. *American Psychologist, 54*(8), 594–604. https://www.apa.org /science/leadership/bsa/statistical/tfsi-followup-report.pdf.
Willmes, Klaus (1996). Neyman-Pearson-Theorie statistischen Testens. In: Edgar Erdfelder et al. (Eds.), *Handbuch Quantitative Methoden* (Kap. I, pp. 109–122). Weinheim: Beltz: PVU.
Wipfler, Holger (2017). *Welchen Einfluss hat eine chiropraktische Justierung der oberen Halswirbelsäule mit High Velocity Low Amplitude Techniken auf die Stressregulierung bei symptomfreien Patienten, gemessen an der*

Herzratenvariabilität? Master Thesis im Universitätslehrgang Chiropraktik. Krems: Gesundheitswissenschaften und Biomedizin, Donau-Universität Krems.

Wow, C.W., A. Krishnan & T. D. Wager (2014). Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. *NeuroImage, 91*, 412–419.

Wrinch, D. & Harold Jeffreys (1921). On certain fundamental principles of scientific inquiry. *The London, Edinburgh, and Dublin Philosophical Magazineand Journalof Science, 42*(249), 369–390.

Wrinch, D. & Harold Jeffreys (1923). On certain fundamental principles of scientific inquiry (second paper). *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 45*(266), 368–374.

WTFDOTA (2017). *Can someone explain to me when I divide by n-I as opposed to dividing by n when calculating variance?* [archived blogpost thread]. url: https://www.reddit.com/r/ statistics/comments/4zmiqd-/can_someone_explain_to_me_when_i_divide_ by_n1_as.

Wu, Margaret (Juli 2009). *Issues in Large-scale Assessments*. Keynote address presented at PROMS 200!), July 28-30, 2009), Hong Kong. Techn. Ber. url: http://www.edmeasurement.com. au/_publications/margaret-/Issues_in_large_scale_assessments.pdf.

Yakoni, Tal (10. Jan. 2011). *The psychology of parapsychology, or why good researchers publishing good articles in good journals can still get it totally wrong.* url: http://www.talyarkoni.org/ blog/2011/01/10/the-psychology-of-parapsychology-or-why-good-researchers-publishing- good- articles- in- good- journals- can- still- get-it-totally-wrong.

Yang, C. Yuan (2016-02). *Multiple Imputation for Missing Data: Concepts and New Development*. P267-25. https : / / stats . idre . ucla . edu / wp - content / uploads / 2016 / 02 / multipleimputation.pdf. Rockville, MD.

Yoccoz, Nigel G. (1991). Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bulletin of the Ecological Society of America, 72*(2), 106–111. https://www.researchgate.net/profile/Alessandro_Giuliani/post/How_do_I_objectively_choose_a_statistical_approach_when_I_want_to_publish/attachment/59d6246ec49f478072e99d7f / AS : 272141411389448 @ 1441894976405/download/pvalues.pdf.

Yu, Chong Ho (2003). Resampling methods: concepts, applications, and justification. *Practical Assessment, Research and Evaluation, 8*(19). url: http://PAREonline.net/getvn. asp?v = 8&n = 19 (visited on 09. 01. 2005).

Yuan, Ke-Hai & Scott Maxwell (2005). On the Post Hoc Power in Testing Meaning Differences. *Journal of Educational and Behavioral Statistics, 30*(2), 141–167.

Yuan Tang, Masaaki Horikoshi & Wenxuan Li (2016). ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages. *The R Journal, 8*(2), S. 478–489. https :// journal .r- project.org/archive/2016-2/tang-horikoshi-li.pdf.

Zabalza, M. A. (1991). *Los diarios de clase. Documento para estudiar cualitativamente los dilemas prácticos de los profesores*. Diss. Barcelona: Promociones y Publicaciones Universitarias, S.A.

Zellner, Anton & A. Siow (1980). Posterior odds ratios for selected regression hypotheses. *Trabajos de Estadistica y de Investigacion Operativa, 31*(1), 585–603.