Daniel Peña

FUNDAMENTOS DE ESTADÍSTICA



Alianza Editorial

Daniel Peña

Fundamentos de Estadística

Alianza Editorial

Reservados todos los derechos. El contenido de esta obra está protegido por la Ley, que establece penas de prisión y/o multas, además de las correspondientes indemnizaciones por daños y perjuicios, para quienes reprodujeren, plagiaren, distribuyeren o comunicaren públicamente, en todo o en parte, una obra literaria, artística o científica, o su transformación, interpretación o ejecución artística fijada en cualquier tipo de soporte o comunicada a través de cualquier medio, sin la preceptiva autorización.

Edición electrónica, 2014
www.alianzaeditorial.es

© Daniel Peña Sánchez de Rivera, 2001
© Alianza Editorial, S. A. Madrid, 2014
Juan Ignacio Luca de Tena, 15. 28027 Madrid
ISBN: 978-84-206-8877-0
Edición en versión digital 2014

A Mely, Jorge y Álvaro

Índice

Prólogo 17

1. Introducción 21
   - 1.1 La estadística como ciencia 21
   - 1.2 Algunos problemas que resuelve la estadística 22
   - 1.3 El método estadístico 24
     - 1.3.1 Planteamiento del problema 25
     - 1.3.2 Construcción de un modelo estadístico 26
     - 1.3.3 Recogida de la información muestral 30
     - 1.3.4 Depuración de la muestra 31
     - 1.3.5 Estimación de los parámetros 31
     - 1.3.6 Contrastes de simplificación 31
     - 1.3.7 Crítica y diagnosis del modelo 32
   - 1.4 Notas sobre la historia de la estadística 32
     - 1.4.1 El cálculo de probabilidades 32
     - 1.4.2 La estadística hasta el siglo XIX 37
     - 1.4.3 El nacimiento de la estadística actual 48
     - 1.4.4 La expansión de la estadística durante el siglo XX 41
   - 1.5 Lecturas recomendadas 43

Primera parte
Datos

2. La descripción de una variable 47
   - 2.1 Datos y distribuciones de frecuencias 47

2.1.1 Distribuciones de frecuencias 48
2.1.2 Diagramas de tallo y hojas 49
2.2 Representaciones gráficas 50
2.2.1 Diagrama de Pareto 50
2.2.2 Diagrama de barras 51
2.2.3 Histogramas 53
2.2.4 Gráficos temporales 55
2.2.5 Otras representaciones gráficas 57
2.3 Medidas de centralización y dispersión 59
2.3.1 Medidas de centralización 59
2.3.2 Medidas de dispersión 62
2.4 Medidas de asimetría y curtosis 66
2.4.1 Coeficiente de asimetría 66
2.4.2 Coeficiente de curtosis 67
2.4.3 Otras medidas características 70
2.5 Datos atípicos y diagramas de caja 72
2.5.1 Datos atípicos 72
2.5.2 Diagrama de caja 73
2.6 Transformaciones 77
2.6.1 Transformaciones lineales 77
2.6.2 Transformaciones no lineales 78
2.7 Resumen del capítulo y consejos de cálculo 86
2.8 Lecturas recomendadas 87

3. Descripción conjunta de varias variables 89
3.1 Distribuciones de frecuencias multivariantes 89
3.1.1 Distribución conjunta 90
3.1.2 Distribuciones marginales 91
3.1.3 Distribuciones condicionadas 92
3.1.4 Representaciones gráficas 94
3.2 Medidas de dependencia lineal 96
3.2.1 Covarianza 96
3.2.2 Correlación 97
3.3 Recta de regresión 98
3.3.1 Correlación y regresión 101
3.4 Vector de medias 102
3.5 Matriz de varianzas y covarianzas 103
3.5.1 Varianza efectiva 104
3.6 Resumen del capítulo y consejos de cálculo 110
3.7 Lecturas recomendadas 110
Apéndice 3A: Números índice 111
Apéndice 3B: Análisis descriptivo de series 112
Apéndice 3C: La presentación de datos en tablas 113
Apéndice 3D: Propiedades de la matriz de covarianzas 115

Segunda parte
Modelos

4. Probabilidad y variables aleatorias 121
   4.1 Introducción 121
   4.2 Probabilidad y sus propiedades 122
      4.2.1 Concepto 122
      4.2.2 Definición y propiedades 124
      4.2.3 La estimación de probabilidades en la práctica 126
   4.3 Probabilidad condicionada 128
      4.3.1 Concepto 128
      4.3.2 Independencia de sucesos 131
      4.3.3 Teorema de Bayes 133
   4.4 Variables aleatorias 140
      4.4.1 Variables aleatorias discretas 140
      4.4.2 Variables aleatorias continuas 142
      4.4.3 Medidas características de una variable aleatoria 147
      4.4.4 Transformaciones 151
   4.5 Resumen del capítulo 159
   4.6 Lecturas recomendadas 160
      Apéndice 4A: Álgebras de probabilidad 161
      Apéndice 4B: Cambio de variable en el caso general 164

5. Modelos univariantes de distribución de probabilidad 165
   5.1 El proceso de Bernoulli y sus distribuciones asociadas 165
      5.1.1 Proceso de Bernoulli 165
      5.1.2 Distribución de Bernoulli 166
      5.1.3 Distribución binomial 167
      5.1.4 Distribución geométrica 168
   5.2 El proceso de Poisson y sus distribuciones asociadas 171
      5.2.1 El proceso de Poisson 171
      5.2.2 La distribución de Poisson 172
      5.2.3 Distribución exponencial 174
   5.3 Distribuciones de duraciones de vida 177
   5.4 La distribución normal 181
   5.5 La normal como aproximación de otras distribuciones 184
      5.5.1 El teorema central del límite 184
      5.5.2 Relación entre binomial, Poisson y normal 186
   5.6 La distribución lognormal 189
   5.7 Deducción de distribuciones: el método de Montecarlo 193
      5.7.1 Introducción 193
      5.7.2 El método de Montecarlo 195
      5.7.3 Aplicaciones 198
   5.8 Distribuciones deducidas de la normal 201
      5.8.1 La distribución χ^2 de Pearson 201
      5.8.2 La distribución *t* de Student 202
      5.8.3 La distribución *F* de Fisher 204
   5.9 Distribuciones mezcladas 204

5.10 Resumen del capítulo y consejos de cálculo 207
5.11 Lecturas recomendadas 210
Apéndice 5A: Función generatriz de momentos 210
Apéndice 5B: Distribución hipergeométrica 213
Apéndice 5C: Distribución gamma 214
Apéndice 5D: Distribución beta 215

6. Modelos multivariantes 217
6.1 Variables aleatorias vectoriales 217
6.1.1 Concepto 217
6.1.2 Distribución conjunta 218
6.1.3 Distribuciones marginales 219
6.1.4 Distribuciones condicionadas 222
6.1.5 Teorema de Bayes 224
6.2 Independencia entre variables aleatorias 225
6.3 Esperanzas de vectores aleatorios 229
6.3.1 Concepto 229
6.3.2 Esperanza de sumas y productos 229
6.4 Covarianzas y correlaciones 230
6.4.1 Covarianza 230
6.4.2 Correlación 231
6.4.3 Varianza de sumas y diferencias 231
6.4.4 Matriz de varianzas y covarianzas 232
6.5 Esperanzas y varianzas condicionadas 234
6.5.1 Esperanzas condicionadas 234
6.5.2 Varianzas condicionadas 236
6.6 Transformaciones de vectores aleatorios 237
6.6.1 Concepto 237
6.6.2 Esperanzas de transformaciones lineales 238
6.7 La distribución multinomial 239
6.8 La normal *n*-dimensional 242
6.9 Resumen del capítulo y consejos de cálculo 249
6.10 Lecturas recomendadas 250
Apéndice 6A: El concepto de distancia y sus aplicaciones 250

Tercera parte
Inferencia

7. Estimación puntual 257
7.1 Introducción a la inferencia estadística 257
7.2 Métodos de muestreo 260
7.2.1 Muestra y población 260
7.2.2 Muestreo aleatorio simple 260
7.2.3 Otros tipos de muestreo 261
7.3 La estimación puntual 265
7.3.1 Fundamentos 265
7.3.2 La identificación del modelo 266
7.3.3 El método de los momentos 269

7.4 La distribución de un estimador en el muestreo 270
7.4.1 Concepto 270
7.4.2 Distribución en el muestreo de una proporción 271
7.4.3 Distribución muestral de la media 272
7.4.4 Distribución muestral de la varianza. Caso general 273
7.4.5 Distribución muestral de la varianza en poblaciones normales. 276
7.5 Propiedades de los estimadores 281
7.5.1 Centrado o insesgado 281
7.5.2 Eficiencia o precisión 283
7.5.3 Error cuadrático medio 285
7.5.4 Consistencia 287
7.5.5 Robustez 287
7.5.6 Punto de ruptura de un estimador 289
7.5.7 Propiedades de los estimadores por momentos 291
7.6 Estimadores de máxima verosimilitud 292
7.6.1 Introducción 292
7.6.2 La distribución conjunta de la muestra 292
7.6.3 La función de verosimilitud 295
7.6.4 Estadísticos suficientes 301
7.6.5 El método de máxima verosimilitud 303
7.6.6 Propiedades de los estimadores máximo-verosímiles 305
7.7 Resumen del capítulo y consejos de cálculo 311
7.8 Lecturas recomendadas 312
Apéndice 7A: Muestreo en poblaciones finitas 312
Apéndice 7B: Estimadores eficientes, el concepto de información 313

8. Estimación por intervalos 319
8.1 Introducción 319
8.2 Metodología 321
8.2.1 La selección del estadístico pivote 321
8.2.2 La determinación de los límites 322
8.3 Intervalos para medias de poblaciones normales 323
8.3.1 Varianza conocida 323
8.3.2 Varianza desconocida 325
8.4 Intervalo para medias. Caso general 326
8.4.1 Proporciones 327
8.5 Intervalo para varianzas de poblaciones normales 327
8.6 Intervalo para la diferencia de medias, poblaciones normales 330
8.6.1 Caso de varianzas iguales 330
8.6.2 Caso de varianzas desiguales 331
8.7 Diferencias de medias. Caso general 332
8.8 Intervalo para la razón de varianzas en poblaciones normales 333
8.9 Intervalos asintóticos 336
8.10 Determinación del tamaño muestral 338
8.11 La estimación autosuficiente de intervalos de confianza *(bootstrap)* 340
8.11.1 Introducción 340
8.11.2 La estimación autosuficiente *(bootstrap)* 341
8.12 Resumen del capítulo y consejos de cálculo 348

8.13 Lecturas recomendadas 350
Apéndice 8A: El método herramental *(jackknife)* 350
Apéndice 8B: Construcción mediante ordenador de intervalos de confianza por el método autosuficiente 352

9. Estimación bayesiana 357
9.1 Introducción 357
9.2 Distribuciones a priori 360
9.2.1 Distribuciones conjugadas 362
9.2.2 Distribuciones de referencia 364
9.3 Estimación puntual 365
9.4 Estimación de una proporción 366
9.5 Estimación de la media en poblaciones normales 369
9.6 Comparación con los métodos clásicos 372
9.7 Resumen del capítulo y consejos de cálculo 374
9.8 Lecturas recomendadas 375

10. Contraste de hipótesis 377
10.1 Introducción 377
10.2 Tipos de hipótesis 380
10.2.1 Hipótesis nula 380
10.2.2 Hipótesis alternativa 381
10.3 Metodología del contraste 382
10.3.1 Medidas de discrepancia 383
10.3.2 Nivel de significación y región de rechazo 383
10.3.3 El nivel crítico *p* 386
10.3.4 Potencia de un contraste 387
10.4 Contrastes para una población 391
10.4.1 Contraste para una proporción 391
10.4.2 Contraste de la media 393
10.4.3 Contraste de varianzas, poblaciones normales 395
10.5 Comparación de dos poblaciones 397
10.5.1 Comparación de dos proporciones 397
10.5.2 Comparación de medias, varianzas iguales, muestras independientes 399
10.5.3 Comparación de medias, muestras dependientes apareadas 400
10.5.4 Comparación de varianzas 402
10.5.5 Comparación de medias, muestras independientes, varianzas distintas 404
10.6 Interpretación de un contraste de hipótesis 409
10.6.1 Intervalos y contrastes 409
10.6.2 Resultados significativos y no significativos 410
10.7 Contrastes de la razón de verosimilitudes 411
10.7.1 Introducción 411
10.7.2 Contraste de hipótesis simple frente alternativa simple 411
10.7.3 Contrastes de hipótesis compuestas 413
10.7.4 Contrastes para varios parámetros 416
10.8 Resumen del capítulo 425

10.9 Lecturas recomendadas 425
Apéndice 10A: Deducción del contraste de verosimilitudes 427
Apéndice 10B: Test de razón de verosimilitudes y test de multiplicadores de Lagrange 428

11. Decisiones en incertidumbre 431
11.1 Introducción 431
11.2 Costes de oportunidad 432
11.3 El valor de la información 434
11.4 Decisiones con información muestral 436
11.4.1 El valor de la muestra 436
11.5 Utilidad 443
11.5.1 El criterio del valor esperado 443
11.5.2 El riesgómetro 444
11.5.3 La función de utilidad 446
11.6 La curva de utilidad monetaria 449
11.7 Inferencia y decisión 454
11.7.1 Estimación y decisión 454
11.7.2 Contrastes y decisiones 456
11.8 Resumen del capítulo 458
11.9 Lecturas recomendadas 458

12. Diagnosis y crítica del modelo 459
12.1 Introducción 459
12.2 La hipótesis sobre la distribución 460
12.2.1 Efecto de un modelo distinto del supuesto 460
12.2.2 El contraste χ^2 de Pearson 461
12.2.3 El contraste de Kolmogorov-Smirnov 466
12.2.4 Contrastes de normalidad 469
12.2.5 Soluciones 476
12.2.6 Transformaciones para conseguir la normalidad 477
12.2.7 Estimación no paramétrica de densidades 488
12.3 La hipótesis de independencia 493
12.3.1 Dependencia y sus consecuencias 493
12.3.2 Identificación 495
12.3.3 Contraste de rachas 495
12.3.4 Contraste de autocorrelación 497
12.3.5 Tratamiento de la dependencia 501
12.4 La homogeneidad de la muestra ..... ‘ 501
12.4.1 Heterogeneidad y sus consecuencias 501
12.4.2 Poblaciones heterogéneas: la paradoja de Simpson 502
12.4.3 Identificación de la heterogeneidad: contraste de Wilcoxon 504
12.4.4 Análisis de tablas de contingencia 508
12.4.5 El efecto de datos atípicos 514
12.4.6 Test de valores atípicos 516
12.4.7 Tratamiento de los atípicos 517
12.5 Resumen del capítulo 518
12.6 Lecturas recomendadas 518

Apéndice 12A: El contraste χ^2 de Pearson 521
Apéndice 12B: Deducción del contraste de Shapiro y Wilk 523
Apéndice 12C: Selección gráfica de la transformación 525
Apéndice 12D: Estimadores robustos iterativos 526

Cuarta parte

Control de calidad

13. Control de calidad 535
13.1 Introducción 535
13.1.1 Historia del control de calidad 536
13.1.2 Clasificación de los sistemas de control 537
13.2 Fundamentos del control de procesos 538
13.2.1 El concepto de proceso bajo control 538
13.2.2 Gráficos de control 540
13.3 El control de procesos por variables 541
13.3.1 Introducción 541
13.3.2 Determinación de la variabilidad del proceso 541
13.4 Gráficos de control por variables 542
13.4.1 Gráfico de control para medias 542
13.4.2 Gráfico de control para desviaciones típicas 545
13.4.3 Gráfico de control para rangos 547
13.4.4 Estimación de las características del proceso 549
13.5 Implantación del control por variables 551
13.5.1 Eficacia del gráfico de la media 552
13.5.2 Curva característica de operación 555
13.5.3 Interpretación de gráficos de control 557
13.6 Intervalos de tolerancia 560
13.6.1 La función de costes para el cliente 560
13.6.2 La determinación de tolerancias justas para el cliente 562
13.6.3 El coste de no calidad 563
13.7 El concepto de capacidad y su importancia 564
13.7.1 Índice de capacidad 564
13.7.2 Un indicador alternativo de capacidad 567
13.8 El control de fabricación por atributos 570
13.8.1 Fundamentos 570
13.8.2 El estudio de capacidad 570
13.8.3 Gráficos de control 573
13.9 El control de fabricación por números de defectos 574
13.9.1 Fundamentos 574
13.9.2 Estudios de capacidad y gráficos de control 575
13.10 Los gráficos de control como herramientas de mejora del proceso 577
13.10.1 La mejora de procesos 577
13.10.2 El enfoque seis sigma 578
13.11 El control de recepción 581
13.11.1 Planteamiento del problema 581
13.11.2 El control simple por atributos 582
13.11.3 Planes de muestreo 585

13.11.4 Plan japonés JIS Z 9002 585
13.11.5 Plan Military-Standard (MIL-STD-105D; ISO 2859; UNE 66020) 585
13.11.6 Planes de control rectificativo: Dodge-Romig 597
13.12 Resumen del capítulo 601
13.13 Lecturas recomendadas 602
Apéndice 13A: Cálculo de gráficos de control 603

Tablas:

Explicación de las tablas 607
Tabla 1: Números aleatorios 613
Tabla 2: Probabilidades binomiales acumuladas 615
Tabla 3: Probabilidades de Poisson acumuladas 617
Tabla 4: Distribución normal estandarizada, $N(0,1)$ 618
Tabla 5: Distribución *t* de Student 619
Tabla 6: Distribución *chi*-cuadrado de Pearson 620
Tabla 7: Distribución *F* 621
Tabla 8: Contraste de Kolmogorov-Smirnov 623
Tabla 9: Contraste de Kolmogorov-Smirnov (Lilliefors) 624
Tabla 10: Coeficientes del contraste de Shapiro-Wilk 625
Tabla 11: Percentiles del estadístico *W* de Shapiro y Wilk 627
Tabla 12: Test de rachas 629
Tabla 13: Papel probabilístico normal 631

Formulario 633

Resolución de ejercicios 643

Bibliografía 665

Índice analítico 675

Prólogo

Este libro es el resultado de veinticinco años de experiencia explicando estadística a estudiantes de ingeniería, economía y administración de empresas y otras licenciaturas universitarias. Cubre los conocimientos básicos que estos profesionales deben adquirir como herramientas imprescindibles para su trabajo y como parte de una formación necesaria para entender la ciencia moderna y evaluar la información cuantitativa que como ciudadanos reciben en un mundo donde la estadística juega un papel creciente.

El libro se estructura siguiendo las etapas de construcción de un modelo estadístico. Tras un capítulo introductorio que presenta el contenido global del libro y una breve introducción histórica a los métodos estudiados, los siguientes capítulos siguen la secuencia de una investigación estadística: análisis exploratorio inicial de los datos disponibles (primera parte, datos, capítulos 2 y 3), construcción de un modelo probabilístico (segunda parte, capítulos 4, 5 y 6) y ajuste del modelo a los datos (tercera parte, inferencia, capítulos 7, 8, 9, 10 y 11). Como aplicación de estas ideas, se presenta en la cuarta parte un capítulo de control de calidad, dirigido especialmente a estudiantes que vayan a trabajar en el mundo empresarial, aunque los conceptos y métodos que se exponen son igualmente útiles para mejorar el funcionamiento de cualquier organización.

Este libro está concebido como texto para un primer curso cuatrimestral de estadística orientado a sus aplicaciones. Por esta razón se incluyen temas de gran importancia práctica que no aparecen habitualmente en libros de texto básicos, como la familia Box-Cox de transformaciones, el concepto

de varianza promedio, las relaciones entre los modelos básicos de distribución de probabilidad, las distribuciones mezcladas, el estudio detallado del método de máxima verosimilitud, el concepto de métodos robustos, la combinación de estimadores, la estimación bayesiana, los métodos autosuficientes *(bootstrap)*, los métodos no paramétricos de estimación de densidades, el análisis de homogeneidad de una muestra, el estudio de datos atípicos y la función de autocorrelación muestral. Estas ideas deben introducirse desde el principio porque, de acuerdo con mi experiencia, el estudiante va a necesitarlas en sus primeros análisis estadísticos con datos reales.

A lo largo del libro se ha pretendido ilustrar los conceptos teóricos con ejemplos y, para reforzar y contrastar su asimilación, se han incluido numerosos ejercicios y problemas cuyas soluciones se encuentran al final del volumen. Estos ejercicios se conciben como parte importante del aprendizaje del estudiante y, por tanto, ciertos conceptos teóricos se complementan o generalizan en ellos.

Es tan incompleto estudiar medicina sin ver jamás a un enfermo como estadística sin analizar datos reales. Por otro lado, el análisis de datos hoy es impensable sin utilizar un ordenador. Los ejemplos y análisis de este libro se han realizado con varios programas informáticos, incluyendo Statgraphics, Excel, Minitab, SPSS, S-Plus y Matlab. Cualquiera de estos programas, que se presentan en orden aproximadamente creciente de sofisticación, puede utilizarse para analizar datos estadísticos y es conveniente que el profesor programe las actividades de estudiantes apoyándose en un programa de ordenador que permita explorar las enormes posibilidades del análisis estadístico para comprender realidades complejas y tomar decisiones en incertidumbre. El estudio teórico y la resolución de ejercicios deben completarse con el análisis de problemas reales para que el estudiante compruebe por sí mismo lo que le aporta la teoría estudiada. Por este camino los conceptos teóricos se convierten en herramientas útiles para su futura actividad profesional.

Este libro es una versión revisada del primer tomo de la obra *Estadística: Modelos y Métodos*. La obra se ha revisado, adaptado y reestructurado completamente con tres objetivos. El primero es aprovechar más las posibilidades ofrecidas por la rapidez y simplicidad de los ordenadores actuales. Esto ha llevado en este libro a ampliar la presentación del método de Montecarlo, incluir en el texto con cierto detalle los métodos autosuficientes de estimación *(bootstrap)* mostrando su utilización práctica e introducir numerosos ejercicios y ejemplos que los estudiantes deben resolver utilizando el ordenador. El segundo objetivo es corregir algunos puntos oscuros y mejorar la presentación del material. Esto ha llevado a subdividir los siete capítulos del libro anterior en los trece actuales, a redactar de nuevo muchas secciones, a ampliar la parte de inferencia bayesiana y a reescribir el capítulo de control de calidad. El tercer objetivo es hacer la obra más flexible para distintas audiencias. Por esta razón el segundo tomo de la obra inicial

se ha subdividido en dos libros independientes, *Regresión y diseño de experimentos* y *Análisis de series temporales*, para facilitar su uso como textos en distintos cursos.

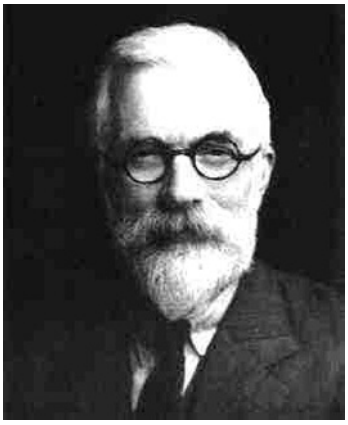
Tengo una deuda especial de gratitud con Rebeca Albacete, María Jesús Sánchez y José Luis Montes, que me han enviado una lista detallada de errores no detectados en ediciones anteriores con excelentes sugerencias de mejora. Gracias a ellos esta edición es más clara y contiene menos erratas. Ángeles Carnero ha conseguido las fotos de estadísticos ilustres buscando en Internet con enorme paciencia y eficacia. Stephan Stigler ha sido de gran ayuda para seleccionar la información histórica. Andrés Alonso, Magdalena Cordero, Pedro Galeano, Miguel Ángel Gómez Villegas, Víctor Guerrero, Jesús Juan, Ana Justel, Agustín Maravall, Francisco Mármol, José Mira, Concepción Molina, Gabriel Palomo, Pilar Poncela, Javier Prieto, Dolores Redondas, Julio Rodríguez, Rosario Romera, Juan Romo, Esther Ruiz, Ismael Sánchez, Santiago Velilla, Teresa Villagarcía, Víctor Yohai y Rubén Zamar han aportado críticas y sugerencias, contribuyendo a mejorar este libro en muchos aspectos. Para todos ellos mi agradecimiento.

Madrid, enero de 2001

En esta nueva edición se han corregido las erratas detectadas y actualizado las referencias. Agradezco mucho la ayuda para llevar a cabo estas mejoras de Adolfo Álvarez, Francisca Blanco, David Casado, Vicente Núñez-Antón, Teresa Villagarcía, Rosario Romera y Henryk Gzyl.

Madrid, junio, 2008

1. Introducción



Ronald Aylmer Fisher (1890-1962)
Científico británico inventor del método de máxima verosimilitud y del diseño estadístico de experimentos. Trabajó en Rothamsted, una estación experimental agrícola en Inglaterra, y fue profesor de eugenesia en la Universidad de Londres. Además de sus numerosas contribuciones a la estadística, que le sitúan como el padre de esta disciplina en el siglo XX, fue un notable genetista, investigador agrario y biólogo.

1.1 La estadística como ciencia

La estadística actual es el resultado de la unión de dos disciplinas que evolucionan independientemente hasta confluir en el siglo XIX: la primera es el cálculo de probabilidades, que nace en el siglo XVII como teoría matemática de los juegos de azar; la segunda es la «estadística» (o ciencia del Estado, del latín *Status*), que estudia la descripción de datos y tiene unas raíces más antiguas. La integración de ambas líneas de pensamiento da lugar a una ciencia que estudia cómo obtener conclusiones de la investigación empírica mediante el uso de modelos matemáticos.

La estadística actúa como disciplina puente entre los modelos matemáticos y los fenómenos reales. Un modelo matemático es una abstracción simplificada de una realidad más compleja, y siempre existirá cierta discrepancia entre lo observado y lo previsto por el modelo. La estadística proporciona

una metodología para evaluar y juzgar estas discrepancias entre la realidad y la teoría. Por tanto, su estudio es básico para todos aquellos que deseen trabajar en ciencia aplicada (sea ésta tecnología, economía o sociología) que requiera el análisis de datos y el diseño de experimentos. La estadística es la «tecnología» del método científico experimental (Mood, 1972).

Además de su papel instrumental, el estudio de la estadística es importante para entender las posibilidades y limitaciones de la investigación experimental, para diferenciar las conclusiones que pueden obtenerse de los datos de las que carecen de base empírica y, en definitiva, para desarrollar un pensamiento crítico y antidogmático ante la realidad.

Muchos ciudadanos ven la estadística con una gran desconfianza: para unos es la ciencia en la que las diferencias individuales quedan ocultas a través de las medias (que se traduce en el dicho popular: «La estadística es la ciencia que explica cómo si tú te comes dos pollos y yo ninguno, nos hemos comido uno cada uno por término medio» y en la famosa frase de Bernard Shaw: «Si un hombre tiene la cabeza en un horno y los pies en una nevera, su cuerpo está a una temperatura media ideal»); para otros es la ciencia mediante la cual con gráficos, tasas de variación y porcentajes se manipula la opinión desde la publicidad, la tecnología o la economía. Vivimos en la era de la estadística y cada aspecto de la actividad humana es medido e interpretado en términos estadísticos.

El único antídoto para esta posible manipulación y para participar efectivamente en la argumentación pública basada en cifras y datos, consustancial a la vida democrática, es comprender el razonamiento estadístico. En este sentido, una formación en los conceptos estadísticos básicos es necesaria para cualquier ciudadano.

1.2 Algunos problemas que resuelve la estadística

Descripción de datos

El primer problema que, históricamente, aborda la estadística es la descripción de datos. Supongamos que se han tomado 1.000 observaciones, que pueden ser gastos de alimentación en una muestra de familias, producción horaria de las máquinas de un taller o preferencias en una muestra de votantes. Se trata de encontrar procedimientos para resumir la información contenida en los datos. Este aspecto se estudia en la primera parte del libro.

Análisis de muestras

Es frecuente que, por razones técnicas o económicas, no sea posible estudiar todos los elementos de una población. Por ejemplo, si para determinar

la resistencia de un elemento es necesario una prueba destructiva, y disponemos de una partida de elementos cuya resistencia se quiere determinar, tendremos que tomar una muestra para no destruir la partida entera. Análogamente, se acude a una muestra para conocer la opinión de la población antes de las elecciones, para estudiar la rentabilidad de un proceso de fabricación o la relación entre el consumo y la renta.

La estadística se utiliza para elegir una muestra representativa y para hacer inferencias respecto a la población a partir de lo observado en la muestra. Éste es el procedimiento aplicado para, por ejemplo:

— Decidir si un proceso industrial funciona o no adecuadamente de acuerdo con las especificaciones.
— Estudiar la relación entre consumo de tabaco y cáncer.
— Juzgar la demanda potencial de un producto mediante un estudio de mercado.
— Orientar la estrategia electoral de un partido político.
— Prever las averías en un taller y diseñar el equipo de mantenimiento.
— Interpretar un test de inteligencia.
— Construir un sistema de reconocimiento de voz.

El análisis de la muestra requiere un modelo probabilístico—cuya construcción será el objeto de la segunda parte de este libro— y la utilización de métodos de inferencia que se expondrán en la tercera parte.

Contrastación de hipótesis

Un objetivo frecuente en la investigación empírica es contrastar una hipótesis. Por ejemplo: ¿Ha mejorado un proceso de fabricación al introducir un elemento nuevo? ¿Es una nueva medicina eficaz para el catarro? ¿Son efectivos el cinturón de seguridad o la limitación de velocidad para reducir las muertes por accidente? ¿Tienen una vida más larga los componentes que tienen el material A que los que no lo tienen? La contrastación de hipótesis requiere una metodología para comparar las predicciones resultantes de la hipótesis con los datos observados y el diseño de un experimento para garantizar que las conclusiones que se extraigan de la experimentación no estén invalidadas por factores no controlados. La metodología estadística para el contraste de hipótesis se expone en el capítulo 10.

Medición de relaciones

Los gastos en alimentación de una familia dependen de sus ingresos, pero es imposible determinar con exactitud cuál será el gasto de una fa-

milia de ingresos dados. Existe entonces una relación no exacta, sino estadística. Determinar y medir estas relaciones es importante porque, debido a los errores de medición, las relaciones que observamos entre variables físicas, sociales o técnicas son, prácticamente siempre, estadísticas.

Preguntas como: ¿Depende la calidad de un producto de los factores A, B y C?, ¿cómo se relaciona el rendimiento escolar con variables familiares y sociológicas?, ¿cuál es la relación entre paro e inflación? tienen que responderse en términos estadísticos. La metodología para analizar estas relaciones se expone en el libro *Regresión y diseño de experimentos,* del mismo autor, que está concebido como extensión de este libro.

Predicción

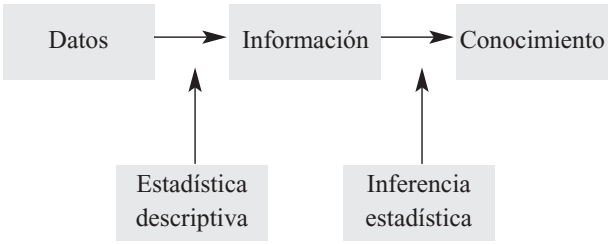
Muchas variables económicas y físicas tienen cierta inercia en su evolución, y aunque sus valores futuros son desconocidos, el estudio de su historia es informativo para prever su evolución futura. Éste es el mecanismo que se utiliza para prever la demanda de un producto, la temperatura en un alto horno o las magnitudes macroeconómicas. La previsión puede mejorarse estudiando la relación entre la variable de interés y otras variables, en el sentido comentado en la sección. Las series temporales se estudian en un texto independiente, concebido como extensión de este libro.

1.3 El método estadístico

El método científico se basa en dos tipos de razonamientos: el deductivo y el inductivo. El método deductivo procede de lo general a lo particular y se utiliza especialmente en el razonamiento matemático: se establecen hipótesis generales que caracterizan un problema y se deducen ciertas propiedades particulares por razonamiento matemático: se establecen hipótesis generales que caracterizan un problema y se deducen ciertas propiedades particulares por razonamientos lógicos. El método inductivo realiza el proceso inverso: a partir de observaciones particulares de ciertos fenómenos se intentan deducir reglas generales.

Una investigación empírica utiliza ambos tipos de razonamiento siguiendo un ciclo deductivo-inductivo: las hipótesis implican propiedades observables en los datos cuyo análisis lleva a formular hipótesis más generales, y así sucesivamente. El método estadístico es el procedimiento mediante el cual se sistematiza y organiza este proceso de aprendizaje iterativo para convertir los datos en información y esta información en conocimiento según el esquema indicado en el cuadro 1.1. La estadística descriptiva se utiliza para sintetizar y resumir los datos transformándolos en información.

Cuadro 1.1 El método estadístico



Esta información es procesada a través de modelos y utilizada para adaptar el modelo a la realidad estudiada, con lo que convertimos la información en conocimiento científico de esa realidad. A continuación se describen las etapas básicas de una investigación estadística.

1.3.1 Planteamiento del problema

Una investigación empírica suele iniciarse con un interrogante del tipo: ¿Cuál es la relación entre...? ¿Qué diferencias existen entre...? ¿Qué ocurriría si...? La primera etapa de la investigación requiere definir el problema en términos precisos, indicando:

a) El ámbito de aplicación, es decir, *la población* que se quiere investigar. Esto exige definir sus límites y caracterizar a sus miembros sin ambigüedad.
b) *Las variables* que debemos observar y cómo medirlas.

Por ejemplo, supongamos que deseamos conocer si la procedencia familiar de un estudiante está relacionada con su rendimiento académico. Tendremos que comenzar definiendo la población que queremos estudiar (por ejemplo, estudiantes matriculados por primera vez en primer curso de una universidad concreta), las variables que definen la procedencia familiar (zona geográfica, estudios de los padres, etc.) y las variables que definen el rendimiento (por ejemplo, nota media en el examen de junio).

Esta fase es fundamental, ya que las conclusiones sólo se aplican a los miembros de la población definida y su validez depende de una selección adecuada de las variables a estudiar.

El resultado de esta fase es una variable respuesta o explicada observable en una o varias poblaciones definidas sin ambigüedad, y un conjunto de variables que podrían explicar esta variable respuesta y que llamaremos variables explicativas.

1.3.2 Construcción de un modelo estadístico

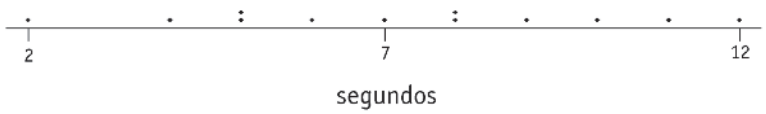
Los modelos estadísticos pueden clasificarse en función de la información que utilizan y del objetivo que pretenden. Cuando la información utilizada corresponde a una única variable, se denominan *modelos univariantes,* cuando incluye además los valores de una o más variables explicativas, se denominan *modelos explicativos.* Por otro lado, si el objetivo es investigar las variables en un instante temporal dado, se denominan *estáticos* o de corte *transversal* (por ejemplo, la relación entre renta y ahorro de las familias españolas en el año 2000), mientras que cuando se desea representar una evolución a lo largo del tiempo se denominan *dinámicos* o *longitudinales.*

En cualquiera de estos cuatro casos, los modelos estadísticos que vamos a estudiar corresponden a una descomposición de los valores de una variable respuesta, *y,* en dos partes. Una parte predecible o sistemática y otra aleatoria, impredecible o residual. El modelo estadístico define la forma de la parte predecible, que representa la respuesta media, y la variabilidad de la impredecible respecto a esa respuesta media. Esta descomposición puede escribirse como:

observación (y) = parte sistemática (predecible)
+ parte aleatoria (impredecible)

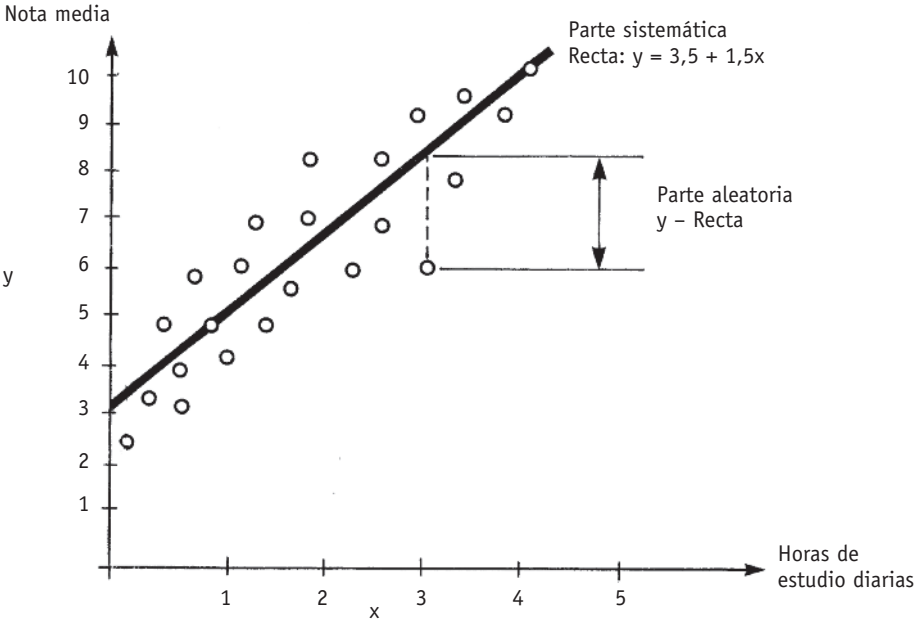
Las figuras 1.1 y 1.2 presentan dos ejemplos de esta descomposición.

Figura 1.1 Tiempo de respuesta en Internet



La primera presenta el tiempo requerido para llegar a una dirección de Internet en doce ocasiones. Cada observación se ha representado por un punto, y la figura muestra que la conexión más rápida se hizo en 2 segundos y la más lenta en 12. Se observa que el tiempo oscila alrededor de un valor central de 7 segundos. Un modelo simple para esta situación es suponer que la conexión se hace en promedio en 7 segundos, pero hay una variabilidad aleatoria en la conexión, de manera que puede tardarse 5 segundos por arriba o por abajo de este valor.

Figura 1.2 Relación entre horas de estudio y nota media



La figura 1.2 representa la relación entre la nota media (variable *y*) obtenida por un grupo de estudiantes en una asignatura y las horas diarias (variable *x*) que en promedio han dedicado a su estudio. Se observa que la nota media depende de las horas de estudio y que los datos se distribuyen alrededor de una recta, que será la parte sistemática o predecible. Esta recta indica un crecimiento lineal de la nota media con el número de horas de estudio. La parte aleatoria será la diferencia entre los valores observados y la recta, y recoge el efecto de todas las variables no consideradas en el modelo (inteligencia de estudiantes, preparación previa, etc.) que producen la variabilidad respecto a la relación promedio.

Estos dos ejemplos son modelos estáticos, ya que estudian la variabilidad en un momento temporal dado. Los modelos de las figuras 1.3 y 1.4 son modelos dinámicos: el primero es *extrapolativo,* ya que utiliza únicamente la información histórica de una serie; el segundo es *explicativo,* ya que introduce otras series como variables explicativas. La figura 1.3 presenta la serie del número de vehículos matriculados cada mes en España en un período de 12 años. La parte sistemática o predecible es ahora mucho más compleja, ya que es la suma de dos componentes:

a) Un componente de tendencia que hace crecer, en promedio, las matriculaciones según una línea recta cuya pendiente varía con el tiempo.

Figura 1.3 Descomposición de la serie de matriculación de vehículos en parte sistemática y parte impredecible o aleatoria

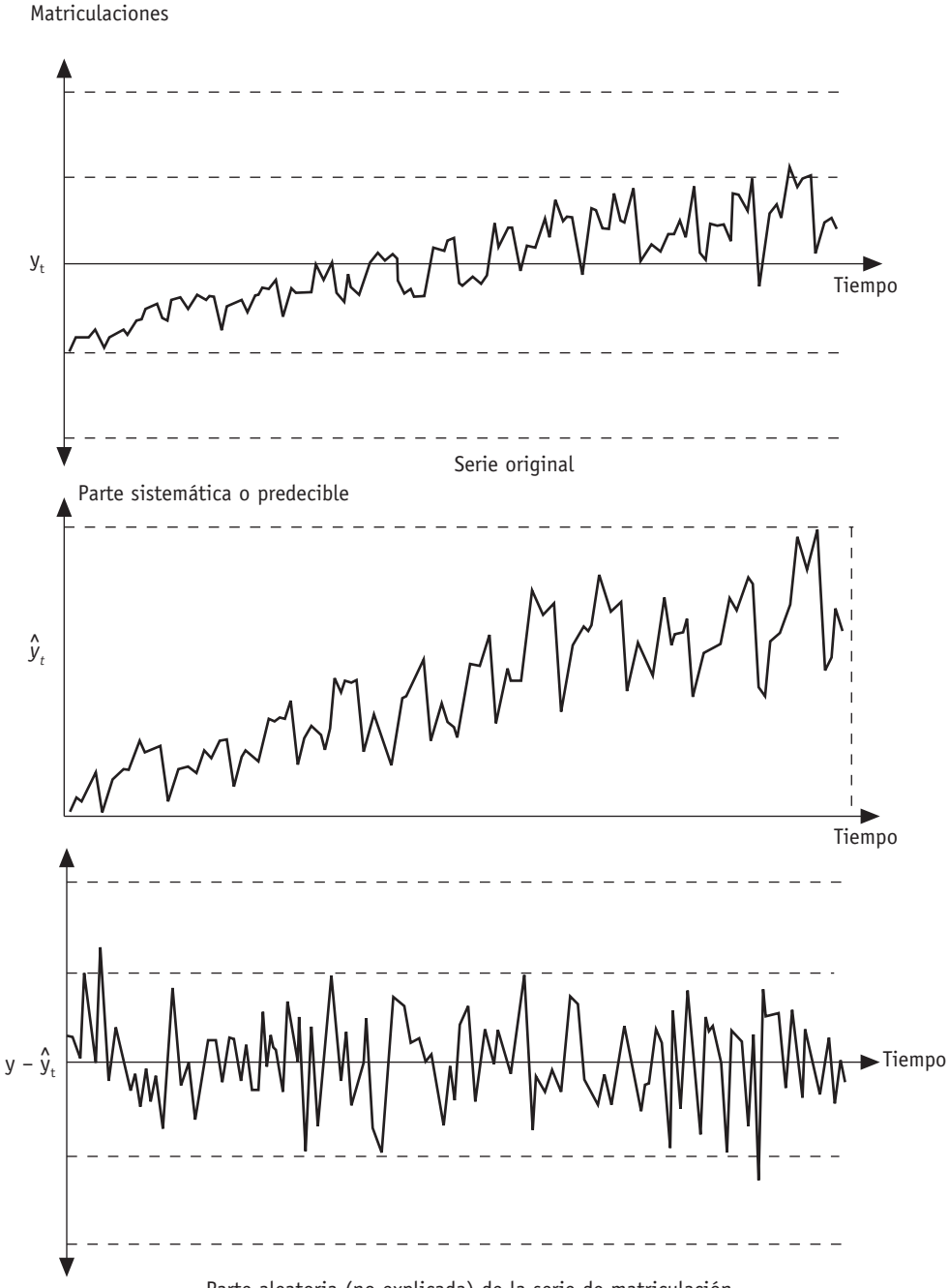
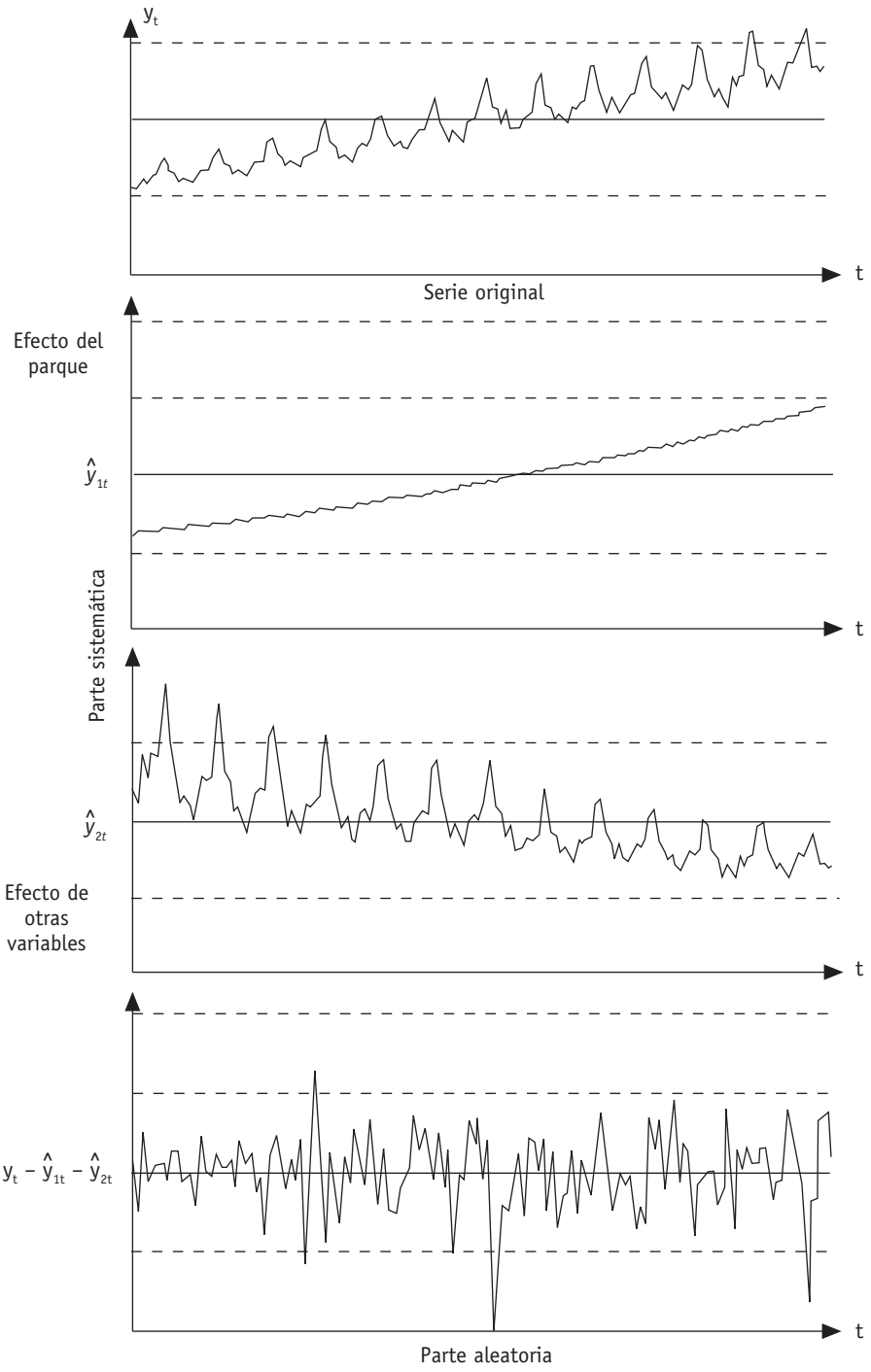


Figura 1.4 Descomposición de la serie de consumo de gasolina



b) Un componente estacional, que hace que cada mes el número de matriculaciones esperadas sea distinto: cada año, descontando la tendencia, se matriculan más vehículos siempre en mayo que en marzo.

Si restamos al número de matriculaciones cada mes, variable y_t, la tendencia y la estacionalidad, cuya suma es la parte predecible, variable $\hat{y}_t$, obtenemos la parte aleatoria o no explicada de la serie que nos proporciona la variabilidad de los datos respecto al valor medio o sistemático de la variable (véase la figura 1.3).

La figura 1.4 presenta un análisis para explicar la evolución de la serie de consumo de gasolina en función del parque de vehículos. La parte sistemática o previsible es ahora la suma de dos componentes. La primera es el efecto debido al aumento del parque de vehículos $\hat{y}_{1t}$, que es una tendencia lineal continuada por el crecimiento del parque. La segunda es la parte sistemática, debida a las otras variables no incluidas en el modelo pero cuya evolución se ha incorporado a la historia de la serie de gasolina. Este efecto es la suma a su vez de una tendencia y de un componente estacional (el consumo de gasolina aumenta en verano) y produce una tendencia decreciente (que puede ser debida al aumento de la eficiencia de los vehículos y a los aumentos del precio de la gasolina) con un efecto estacional superpuesto. Finalmente, la parte aleatoria es la diferencia entre la serie observada y la suma de estos dos componentes explicados, parte explicada o sistemática.

Estos ejemplos muestran las características generales de los modelos estadísticos más frecuentes. Conceptualmente, una variable cualquiera, y, será función de otro gran número de variables, algunas de las cuales pueden no ser observables y cuyo número exacto se desconoce. Un modelo estadístico es una aproximación operativa de esta realidad, que tiene en cuenta explícitamente las variables observables presumiblemente más importantes, y engloba en la parte aleatoria los efectos del resto. Una extensión de estos modelos son los modelos *multivariantes*, donde el interés se centra en un conjunto de variables que se desea explicar conjuntamente. El capítulo 6 presenta una introducción a estos modelos en el caso estático.

1.3.3 Recogida de la información muestral

Una vez construido un modelo del problema, tendremos que medir los valores de las variables de interés. Esta recogida de información puede hacerse de dos formas:

a) Por muestreo.
b) Con un diseño de experimentos.

El muestreo consiste en observar pasivamente una muestra de las variables y anotar sus valores; se utiliza especialmente en modelos extrapolativos.

El diseño de experimentos consiste en fijar los valores de ciertas variables y observar la respuesta de otras. Debe utilizarse siempre que sea posible cuando se desee construir un modelo explicativo. Unicamente tendremos una base empírica sólida para juzgar respecto a relaciones de causalidad entre variables cuando los datos se obtengan mediante un adecuado diseño experimental.

Los fundamentos del muestreo se exponen en el capítulo 7, y los métodos de diseño experimental, en el segundo texto de este trabajo.

1.3.4 Depuración de la muestra

Una regla empírica ampliamente contrastada (Huber, 1984) es esperar entre un 2 y un 5% de observaciones con errores de medición, transcripción, etc. Por tanto, antes de utilizar los datos muestrales conviene aplicar técnicas estadísticas simples, como las que se presentan en el capítulo 2, para identificar valores anómalos y eliminar los errores de medición.

1.3.5 Estimación de los parámetros

Los modelos estadísticos dependen de ciertas constantes desconocidas que llamaremos parámetros. A veces se dispone de información a priori respecto a sus valores, y otras esta información inicial será muy pequeña con relación a la que aportará la muestra. La fase de estimación consiste en utilizar la información disponible para estimar los valores de estos parámetros, así como cuantificar el posible error en la estimación. Los fundamentos de la teoría de la estimación, que son generales para cualquier modelo estadístico, se estudiarán en la tercera parte en los capítulos 7, 8 y 9.

1.3.6 Contrastes de simplificación

Una vez estimados los valores de los parámetros, estudiaremos si el modelo puede simplificarse: por ejemplo, dos parámetros pueden aproximadamente ser iguales, otro puede ser cero, etc. El objetivo de esta fase es conseguir un modelo tan simple como sea posible, es decir, sin más parámetros que los necesarios. Esta fase es especialmente importante en los modelos explicativos, pero aparece en mayor o menor medida en toda investigación estadística. La teoría de contraste de hipótesis se estudiará en el capítulo 10.

1.3.7 Crítica y diagnosis del modelo

Los resultados de las etapas 5 y 6 anteriores se obtienen suponiendo que el modelo es correcto. Esta fase investiga la compatibilidad entre la información empírica y el modelo. De especial interés es comprobar que la parte aleatoria lo es realmente, es decir, no contiene ninguna estructura sistemática. Este aspecto se estudia en el capítulo 12.

Si después de esta fase aceptamos el modelo como correcto, lo utilizaremos para tomar decisiones (capítulo 11) o realizar previsiones de la variable. En caso contrario volveremos a la fase 2 y reformularemos el modelo, repitiendo el proceso hasta conseguir un modelo correcto. Este aspecto cíclico de la investigación se resume en el cuadro 1.2.

La metodología aquí expuesta es iterativa y utiliza tanto el razonamiento deductivo (especialmente en las etapas 2 y 3) como el inductivo (desde la 4 hasta la 7). El cuadro 1.3 presenta dos ejemplos de investigaciones estadísticas: la primera utiliza modelos extrapolativos estáticos y la segunda un modelo explicativo estático.

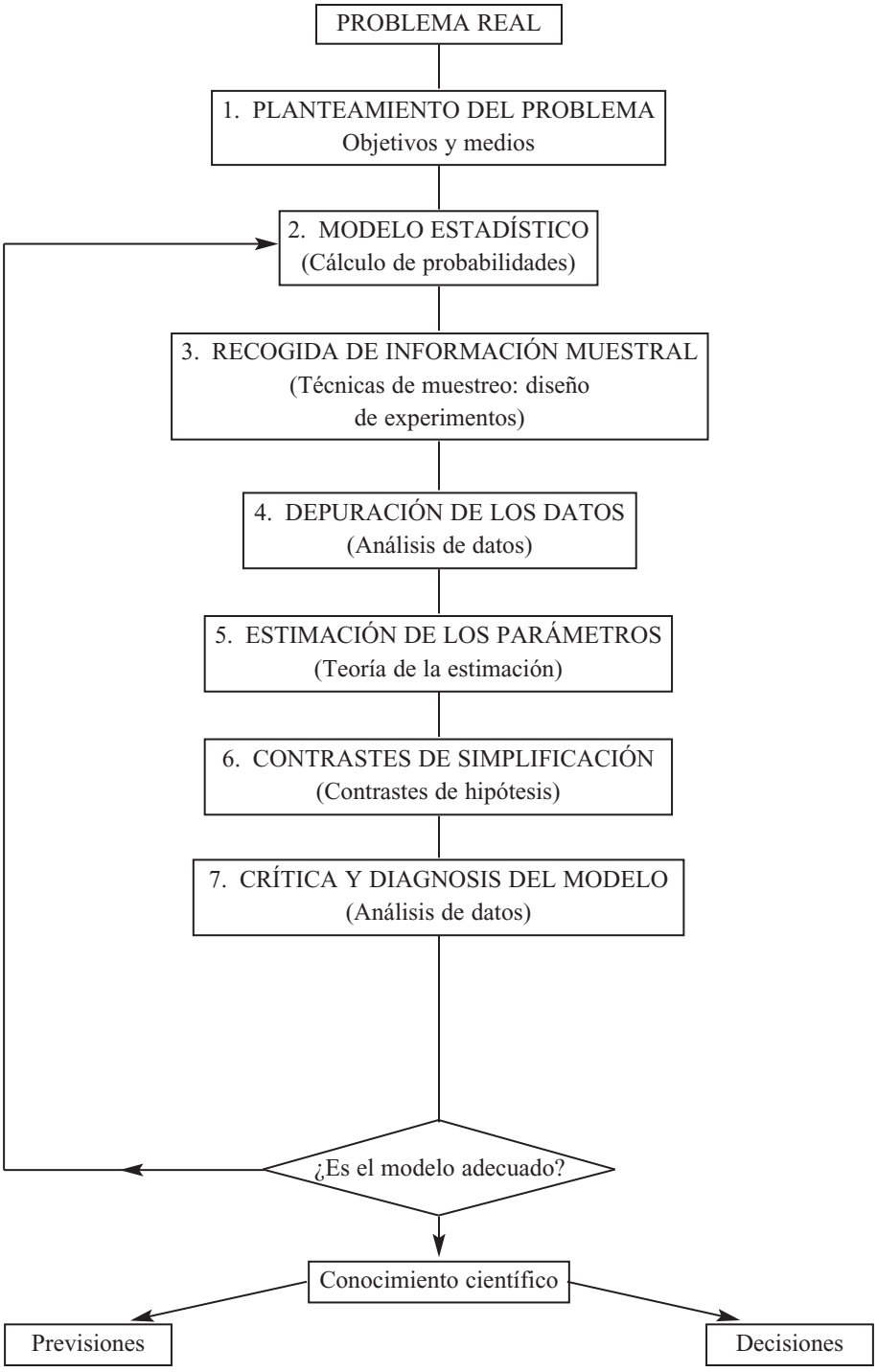
1.4 Notas sobre la historia de la estadística

El conocimiento de la historia de una disciplina es importante, al menos en tres aspectos: el primero, para entender su estado actual de desarrollo y la relación entre sus partes; el segundo, para comprender su terminología, ya que el nombre de una técnica o de un método suele estar asociado a sus orígenes históricos; el tercero, para prever su desarrollo futuro. Por estas razones, consideramos conveniente presentar brevemente algunos rasgos fundamentales de la evolución de la estadística.

1.4.1 El cálculo de probabilidades

La abundante presencia del hueso astrágalo de oveja o ciervo (que constituye el antecedente inmediato del dado) en las excavaciones arqueológicas más antiguas parece confirmar que los juegos de azar tienen una antigüedad de más de 40.000 años, y la utilización del astrágalo en culturas más recientes, Grecia, Egipto y posteriormente Roma, ha sido ampliamente documentada. En las pirámides de Egipto se han encontrado pinturas que muestran juegos de azar que provienen de la primera dinastía (3500 a.C.), y Herodoto se refiere a la popularidad y difusión en su época de los juegos de azar, especialmente mediante la tirada de astrágalos y dados. Los dados más antiguos que se han encontrado se remontan a unos 3.000 años a.C. y se utilizaron tanto en el juego como en ceremonias religiosas.

Cuadro 1.2 Etapas de construcción de un modelo estadístico



Cuadro 1.3 Dos ejemplos de investigaciones estadísticas

Pregunta	¿Cómo diseñar un puesto de servicio?	¿Cómo aumentar el rendimiento de un proceso?
MODELO	Variables: — Número de clientes (x_1) — Tiempo de servicio (x_2) Hipótesis: los clientes • Llegan independientemente. • La probabilidad de llegada de un cliente aumenta exponencialmente con el tiempo. Hipótesis: el tiempo de servicio • Depende de muchos pequeños factores.	Variables: — Rendimiento en % (y) — Temperatura x_1 — Concentración x_2 Hipótesis: • El rendimiento aumenta en promedio linealmente con la temperatura y la concentración. • Para valores fijos de x_1 y x_2 el rendimiento varía aleatoriamente alrededor de su valor medio.
RECOGIDA DE INFORMACIÓN	Muestreo del sistema para estudiar las llegadas de clientes y tiempos de servicio.	Diseño de un experimento en que se varíen x_1 y x_2 y se mida y.
ESTIMACIÓN DE PARÁMETROS	Estimar: • λ, tasa media de llegada. • μ, tiempo medio de servicio. • σ, variabilidad en el tiempo de servicio.	Estimar: • El efecto de la temperatura (b) y el de la concentración (c) sobre el rendimiento. • La variabilidad experimental.
CONTRASTES DE SIMPLIFICACIÓN	¿Tienen todas las semanas la misma λ? ¿Los clientes, el mismo μ y σ?	¿Es el efecto de la temperatura y concentración idéntico ($b = c$)? ¿Puede suponerse $b = 0$?
CRÍTICA DEL MODELO	¿Es cierta la independencia entre llegadas? ¿Son la variabilidad de x_1 y x_2 en la muestra consistentes con las hipótesis?	¿Es la relación entre y (x_1, x_2) lineal? ¿Es la variabilidad de y para x_1, x_2 fijos independiente de los valores concretos de x_1 y x_2?

En las civilizaciones antiguas, el azar se explicaba mediante la voluntad divina. Los oráculos, sacerdotes o pitonisas de Grecia y Roma utilizaban la configuración resultante de tirar cuatro dados para predecir el futuro y revelar la voluntad favorable o desfavorable de los dioses. Por ejemplo, en Grecia clásica y Roma la aparición de la combinación Venus (aparición de 1, 3, 4, 6 al tirar cuatro dados) era favorable, y se ha descubierto en Asia Menor una completa descripción de la interpretación profética de los posibles resultados al tirar cuatro dados. Prácticas similares se han encontrado en culturas tan distantes como la tibetana, la india o la judía.

Como no es posible encontrar una causa o conjunto de causas que permitan predecir el resultado de tirar un dado, las culturas antiguas basadas en el determinismo atribuyeron los resultados de fenómenos aleatorios (dados, presencia de lluvia o fenómenos climáticos, etc.) a la voluntad divina. Piaget ha hecho notar que esta actitud mágica ante el azar se manifiesta igualmente en los niños.

El Renacimiento supuso un nuevo enfoque global de la concepción del mundo, e indujo una observación cualitativamente distinta de muchos fenómenos naturales. En concreto, el abandono progresivo de explicaciones teológicas conduce a una reconsideración de los experimentos aleatorios, y los matemáticos italianos de comienzos del siglo XVI empiezan a interpretar los resultados de experimentos aleatorios simples. Por ejemplo, Cardano, en 1526, establece, por condiciones de simetría, la equiprobabilidad de aparición de las caras de un dado a largo plazo, y Galileo (1564-1642), respondiendo a un jugador que le preguntó por qué es más difícil obtener 9 tirando 3 dados que obtener 10, razonó que de las 216 combinaciones posibles equiprobables 25 conducen a 9 y 27 a 10. Señalamos este dato porque la diferencia empírica entre obtener 9 o 10 es únicamente de $2/216 \simeq 0{,}01$, lo que muestra cómo a finales del siglo XVI existía un intuitivo pero preciso análisis empírico de los resultados aleatorios.

El desarrollo del análisis matemático de los juegos de azar se produce lentamente durante los siglos XVI y XVII, y algunos autores consideran como origen del cálculo de probabilidades la resolución del problema de los puntos en la correspondencia entre Pascal y Fermat en 1654. El problema planteado a estos autores por el caballero de Meré, un jugador empedernido de la Francia del XVII, fue cómo debería repartirse el dinero de las apuestas depositado en la mesa si los jugadores se vieron obligados (presumiblemente por la policía, ya que el juego estaba entonces prohibido) a finalizar la partida sin que existiera un ganador.

El cálculo de probabilidades se consolida como disciplina independiente en el período que transcurre desde la segunda mitad del siglo XVII hasta comienzos del siglo XVIII. En ese período, la teoría se aplica fundamentalmente a los juegos de azar.

Durante el siglo XVIII el cálculo de probabilidades se extiende a problemas físicos y actuariales (seguros marítimos). El factor principal impulsor

de su desarrollo durante este período es el conjunto de problemas de astronomía y física que surgen ligados a la contrastación empírica de la teoría de Newton.

La obra de Newton (1642-1727) constituyó la mayor revolución científica de los siglos XVII y XVIII y su influencia en la evolución de las ciencias físicas es ampliamente conocida. En astronomía, Newton no solamente explicó las leyes de Kepler por el principio de gravitación universal, sino que estableció un modelo global para estudiar las relaciones entre los cuerpos estelares. En física, estableció una teoría común para explicar fenómenos que habían sido objeto de estudios fragmentarios e incompletos como péndulos, planos inclinados, mareas, etc. En matemáticas, contribuyó con Leibnitz a la creación del cálculo diferencial e integral.

Durante el siglo XVIII y parte del XIX la investigación en física y astronomía está dirigida por el paradigma de Newton. Esta investigación se centra en: a) campos de observación y experimentación que la teoría de Newton señala como especialmente relevantes; b) contrastación de las predicciones de la teoría con los datos; c) extender las aplicaciones de la teoría en otros campos. Estas investigaciones van a ser de importancia fundamental en el desarrollo de la estadística.

Un primer problema fue el tratamiento de los errores de medición. Se disponía de varias medidas independientes de una determinada magnitud física y se presentaba el interrogante de cómo combinarlas para obtener un resultado más preciso. Aunque este problema se había planteado en la astronomía desde la antigüedad, la necesidad de comparar con exactitud los datos observados con la teoría requería un tratamiento riguroso del mismo, que va a dar lugar a la teoría de errores.

D. Bernoulli (1700-1782) proporciona la primera solución al problema de estimar una cantidad desconocida a partir de un conjunto de mediciones que, por el error experimental, presentan variabilidad. También desarrolló un test estadístico para determinar si puede aceptarse la hipótesis de que el ordenamiento de las órbitas de los planetas es aleatorio. Este autor fue pionero en la aplicación del cálculo infinitesimal al cálculo de probabilidades.

Pierre Simon, marqués de Laplace (1749-1827), introdujo la primera definición explícita de probabilidad y desarrolló la ley normal como modelo para describir la variabilidad de los errores de medida. También se planteó el problema de predecir una variable conociendo los valores de otras relacionadas con ella y formuló y estimó el primer modelo explicativo estadístico. Es de señalar que, aunque sus procedimientos matemáticos fueron muy «ad hoc», sus resultados fueron sorprendentemente precisos.

La segunda contribución fundamental de este período es debida a Legendre (1752-1833) y Gauss (1777-1855), que resuelven de manera general el problema siguiente de estimación de modelos estáticos: según la teoría, la posición de un planeta en el instante *t,* que llamaremos y_t, es función de

las posiciones de k cuerpos, que representaremos por $x_1, ..., x_k$, y de ciertas constantes desconocidas $\theta_1, ..., \theta_k$. Es decir,

$$y_t = f(\theta_1, ..., \theta_h; x_1, ..., x_k)$$

Disponemos de ciertas observaciones —con cierto error de medida— de las posiciones del planeta y de los cuerpos en cuestión. ¿Cómo determinar las constantes $\theta_1, ..., \theta_h$? ¿Cómo predecir y_t, con la mayor precisión posible dada una observación concreta de valores $x_1, ..., x_k$?

Legendre resolvió estos problemas inventando el método de estimación de mínimos cuadrados, que es todavía hoy la herramienta más utilizada para estimar modelos estadísticos, y Gauss demostró su optimalidad cuando los errores de medida siguen una distribución normal.

Durante la primera mitad del siglo XIX, los matemáticos-astrónomos continúan ampliando la teoría de errores y podemos observar la aparición de problemas y métodos que van a tener gran influencia posterior. Bravais (1846), geólogo y astrónomo, es el primero en considerar la relación entre errores de medida dependientes entre sí, Benjamin Pierce (1852) propone el primer criterio para rechazar observaciones heterogéneas con el resto y S. Newcomb, el más famoso astrónomo americano del XIX, introduce los primeros métodos de estimación cuando hay errores fuertes en algunos datos (estimación robusta).

Por lo tanto, a mediados del siglo XIX existen ya las herramientas básicas que van a dar lugar a la estadística actual. Sin embargo, la aplicación de estos principios va a restringirse a la física y la astronomía, sin ejercer influencia sobre otras áreas de conocimiento.

En particular, estos avances tienen poca influencia sobre una disciplina científica cuyo campo de estudio es el análisis cuantitativo de datos demográficos, sociales y económicos y que se conoce, desde el siglo XVII, con el nombre de estadística.

1.4.2 La estadística hasta el siglo XIX

Desde la antigüedad, los estados han recogido información sobre la población y riqueza que existía en sus dominios. Los censos romanos, los inventarios de Carlomagno de sus posesiones, etc., pueden considerarse precedentes de la institucionalización de la recogida de datos demográficos y económicos por los estados modernos, principalmente por razones fiscales. Esta aritmética política o estadística descriptiva evoluciona durante los siglos XVII y XVIII tomando progresivamente un carácter más cuantitativo.

El primer intento de aplicar un razonamiento propiamente estadístico, en el sentido actual del término, a datos demográficos es debido, en 1662, a Graunt. Este autor se planteó el problema de estimar la población inglesa

de su época y fue capaz, a partir de una muestra, de estimar por primera vez tasas de mortalidad por edades y deducir la frecuencia de nacimientos de hombres y mujeres, entre otros análisis demográficos relevantes. El tipo de razonamiento de Graunt es puramente analítico y desligado completamente del concepto de probabilidad. En la misma línea Petty, en su *Political Arithmetic*, publicado en 1690, analiza datos demográficos, así como datos económicos de ingresos, educación y comercio.

Las primeras tablas completas de mortalidad fueron publicadas por Edmund Halley en 1693, que estudió el problema de los seguros de vida. Durante el siglo XVIII se produce un rápido crecimiento, principalmente en Inglaterra, de los seguros de vida y los seguros marítimos y, debido en gran parte a la influencia de las ideas de Graunt y Petty, se comienzan a realizar los primeros censos oficiales. El primer censo del que se tiene noticias fue realizado por España en Perú en 1548 bajo la dirección del virrey D. Pedro de la Fasca. En Europa, el primer censo se realiza en Irlanda en 1703, y en España, el primero se efectúa en 1787 impulsado por el conde de Floridablanca. A comienzos del siglo XIX puede afirmarse que la casi totalidad de los países europeos recogen información oficial mediante censos de datos demográficos, económicos, climáticos, etc. Paralelamente, surgen las Agencias Oficiales de Estadística y en 1834 se crea en Londres la Royal Statistical Society, seguida, en 1839, por la American Statistical Association.

Durante el siglo XVIII y la mayor parte del siglo XIX, la estadística evoluciona como ciencia separada del cálculo de probabilidades. Aunque A. de Moivre y Deparcieux, entre otros, aplican el cálculo de probabilidades a datos demográficos, y Condorcet y Laplace a problemas de aritmética política, existe durante este período escasa comunicación entre ambas disciplinas. Una contribución importante hacia dicha síntesis es debida a A. Quetelet (1846), que sostuvo la importancia del cálculo de probabilidades para el estudio de datos humanos. Quetelet demostró que la estatura de los reclutas de un reemplazo seguía una distribución normal, e introdujo el concepto de «hombre medio». Sin embargo, la diferencia de concepción y de lenguaje entre los matemáticos-astrónomos y los estadísticos-demógrafos dificultó la interacción entre ambos grupos. La unión entre ambas corrientes va a producirse a comienzos del siglo xx, favorecida, en gran parte, por los nuevos problemas teóricos y metodológicos que planteaba la contrastación empírica de la teoría de Darwin.

1.4.3 El nacimiento de la estadística actual

La revolución que supuso en la física Newton se produjo en la biología por la obra de Darwin. Dos facetas importantes de esta teoría eran: a) permitía establecer predicciones sobre la evolución de poblaciones animales que, en

determinadas condiciones, podían ser contrastadas empíricamente; b) la contrastación debería ser estadística, ya que la unidad que va a sufrir la evolución es la población en su conjunto. Los dos mecanismos de la selección natural, producción de variabilidad y selección mediante lucha por la existencia, tienen un atractivo inmediato desde el punto de vista estadístico. La producción de variabilidad mediante el azar entronca con el cálculo de probabilidades; la selección natural, con el estudio de poblaciones y con la idea de correlación. Aquellos organismos que estén más adaptados sobrevivirán un mayor período de tiempo y dejarán un mayor número de descendientes, por lo que tiene que existir una correlación entre determinadas características genéticas transmisibles y el grado de supervivencia y descendencia de los individuos de una especie.

El primero en resaltar la necesidad de acudir a métodos estadísticos para contrastar la teoría de Darwin fue Francis Galton (1822-1911). Galton, primo de Darwin, fue un hombre de profunda curiosidad intelectual que le llevó a viajar por todo el mundo y a realizar actividades tan diversas como redactar leyes para los hotentotes que gobernaban en el sur de África o realizar fecundas investigaciones en meteorología (a él le debemos el término «anticiclón»). La lectura de la obra de Darwin supuso una transformación radical en la vida de Galton, que, casi a los 40 años, dedica sus esfuerzos al estudio de la herencia humana. Su trabajo principal es *Natural Inheritance*, publicado en 1889 (a la edad de 67 años). Galton estudió exhaustivamente la distribución normal e introdujo el concepto de línea de regresión comparando las estaturas de padres e hijos. Galton encontró que los padres altos tenían, en promedio, hijos altos, pero en promedio más bajos que sus padres, mientras que los padres bajos tenían hijos bajos, pero, en promedio, más altos que sus padres. Este fenómeno de regresión se ha encontrado en muchas características hereditarias, de manera que los descendientes de personas extremas en alguna característica estarán, en promedio, más cerca de la media de la población que sus progenitores, produciendo así un efecto de regresión (vuelta) a la media de la población.

La importancia de Galton radica no solamente en el nuevo enfoque que introduce en el problema de la dependencia estadística, sino también en su influencia directa sobre Weldon, K. Pearson, R. A. Fisher y Edgeworth entre otros. El primer departamento de estadística en el sentido actual de la palabra fue patrocinado por él y llevó su nombre, y la revista *Biométrica* fue posible gracias a su generoso apoyo económico.

El enfoque estadístico propugnado por Galton para el estudio de los problemas de la evolución en *Natural Inheritance* es aceptado entusiásticamente por W. R. F. Weldon (1860-1906), entonces catedrático de zoología en la Universidad de Londres. Weldon abandona el camino de los estudios embriológicos y morfológicos como medio de contrastar las hipótesis de Darwin y comienza a investigar en la aplicación de los métodos estadísticos a la biología animal. En 1893 (Weldon, 1893), escribe:

Es necesario insistir en que el problema de la evolución animal es esencialmente un problema estadístico [...] debemos conocer: a) el porcentaje de animales que exhiben un cierto grado de anormalidad respecto a un carácter; b) el grado de anormalidad de otros órganos que acompaña a las normalidades de uno dado; c) la diferencia entre la tasa de mortalidad en animales con diferentes grados de anormalidad respecto a un órgano; d) la anormalidad de los descendientes en términos de anormalidad de los padres y viceversa.

La resolución de estos problemas requiere el desarrollo de métodos estadísticos más avanzados que los existentes, y Weldon busca para ello la colaboración de un matemático y filósofo: K. Pearson (1857-1936). La colaboración de estos dos autores y el apoyo de Galton van a constituir el impulso generador de la corriente de contribuciones que va a fundamentar la estadística actual.

El lector encontrará en los capítulos siguientes varias de las contribuciones de K. Pearson que llevan su nombre. Para facilitar la aplicación de los nuevos métodos, dados los escasos medios de cálculo disponibles a finales del siglo XIX, Pearson dedicó una parte importante de sus esfuerzos a la publicación de tablas estadísticas que permitieran la utilización práctica de los nuevos métodos, con lo que contribuyó, decisivamente, a su rápida difusión.

El laboratorio de K. Pearson se convierte en un polo de atracción para las personas interesadas en el análisis empírico de datos. W. S. Gosset (1876-1937), que trabajaba en la firma cervecera Guinness de Dublín, fue una de las personas que acudieron a Londres a estudiar bajo el patrocinio de Pearson. Gosset se había encontrado en sus investigaciones sobre los efectos de las características de la materia prima en la calidad de la cerveza final con el problema de las pequeñas muestras. No era posible económicamente, en este caso, obtener las grandes cantidades de datos que permitirían utilizar los métodos para muestras grandes desarrolladas por Pearson y su escuela. Para resolver el problema, Gosset realizó el primer trabajo de investigación estadística mediante el método de Montecarlo, tomando 750 muestras aleatorias de cuatro elementos de los datos recopilados por W. R. McDonnell sobre la estatura y la longitud del dedo corazón de 3.000 delincuentes, con los que simuló el proceso de tomar muestras de una distribución normal y obtuvo la distribución *t*, que publicó con el pseudónimo de Student, ya que Guinness no permitía divulgar las investigaciones de sus empleados.

Los fundamentos de la estadística actual y muchos de los métodos de inferencia expuestos en este libro son debidos a R. A. Fisher (1890-1962). Fisher se interesó primeramente por la eugenesia, lo que le conduce, siguiendo los pasos de Galton, a la investigación estadística. Sus trabajos culminan con la publicación de *Statistical Methods for Research Workers.* En él aparece ya claramente el cuerpo metodológico básico que constituye la

estadística actual: el problema de elegir un modelo a partir de datos empíricos, la deducción matemática de las propiedades del mismo (cálculo de probabilidades), la estimación de los parámetros condicionados a la bondad del modelo y la validación final del mismo mediante un contraste de hipótesis.

1.4.4 La expansión de la estadística durante el siglo xx

Entre 1920 y el final de la Segunda Guerra Mundial se extiende la aplicación de los métodos estadísticos en áreas tan diversas como la ingeniería (control de calidad por Shewart, métodos de predicción y control de procesos y codificación de señales por Wiener y Shannon), la economía (estimación de ecuaciones de oferta y demanda, índices de precios, medición de la riqueza y de la pobreza), la física (teoría cinética de los gases), la antropología (clasificación de restos arqueológicos), la psicología (medición de la inteligencia y teoría de test) o la medicina (pruebas para determinar la eficacia de nuevos tratamientos). La búsqueda de respuestas a los nuevos interrogantes planteados por estas aplicaciones impulsan, a su vez, el desarrollo de nuevos métodos estadísticos. Los problemas en agronomía conducen a Fisher a crear la teoría de diseños experimentales, y un problema de discriminación en antropología (concretamente la clasificación de cráneos; véase J. Box, 1978) lleva a Fisher a inventar el análisis discriminante. El análisis factorial surge ligado a problemas en la psicología, y, en general, la economía y las ciencias sociales impulsan el desarrollo de métodos para medir la relación entre variables (métodos de regresión) y analizar muchas variables conjuntamente (métodos multivariantes). Los problemas de ingeniería conducen a un estudio sistemático de la teoría de modelos dinámicos (procesos estocásticos) y a la creación de la teoría de predicción y de extracción de señales de Wiener y Kolmogorov en los años cuarenta. Las necesidades en el control de procesos sugieren a E. S. Pearson la creación de la teoría general de contraste de hipótesis, conjuntamente con J. Neyman, y el trabajo en aplicaciones industriales de la estadística del Statistical Research Group en Columbia durante la Segunda Guerra Mundial condujo a Wald a inventar los contrastes secuenciales para el control de recepción, punto de partida básico en el nacimiento y desarrollo de la teoría estadística de decisión.

A partir de 1950 podemos considerar que comienza la época moderna de la estadística. Algunos de los aspectos diferenciales respecto a los períodos anteriores son:

a) La aparición del ordenador digital, que va a revolucionar la metodología estadística y abrir enormes posibilidades para la construcción de modelos más complejos.

b) El cambio de énfasis en la metodología estadística. La influencia de Neyman, Pearson y Wald en los años cuarenta y cincuenta concentra la investigación teórica en la búsqueda de procedimientos óptimos de estimación y contraste de hipótesis en problemas simplificados. Estos procedimientos óptimos parten de dos premisas fundamentales: a) Los datos de que disponemos han sido generados por un modelo de distribución de probabilidad que es conocido salvo por un vector de parámetros; b) el modelo pertenece a una familia restringida de distribuciones de probabilidad, matemáticamente tratable. Este análisis considera tangencialmente dos problemas centrales del análisis estadístico: a) la identificación de la estructura del modelo y b) los contrastes diagnósticos para, una vez estimado el modelo, decidir si puede rechazarse su estructura básica mediante los datos empíricos.
El enfoque de «métodos óptimos» se apoya en el postulado de continuidad: una pequeña desviación en las hipótesis producirá una pequeña variación en el resultado final. La falsedad de este principio en muchos problemas estadísticos relevantes ha conducido a que actualmente la metodología estadística ponga el énfasis en el proceso iterativo de aprendizaje a partir de los datos en lugar de en la aplicación de un determinado procedimiento óptimo. Con esto, las fases de exploración de los datos, de consideración de modelos flexibles y de procedimientos robustos y generales de estimación han pasado a ocupar el centro de la metodología moderna.

A finales del siglo XX el espectacular aumento de la capacidad de cálculo de los ordenadores y la caída de los costes de almacenamiento de la información han hecho posible la recogida automática de grandes masas de datos en cualquier actividad humana. Por ejemplo, los datos que los satélites nos envían en un solo día bloquearían la capacidad de análisis de un ordenador de los años setenta; un banco, un supermercado o una tienda virtual en Internet adquiere, a través de las operaciones por tarjeta de crédito de sus clientes, bancos de datos del comportamiento de los consumidores que serían intratables hace pocos años, y los procesos de fabricación automática proporcionan información constante de muchas variables de control que, además, evolucionan dinámicamente en el tiempo. El reto más importante de la estadística en este siglo XXI es cómo extraer la información en estas grandes masas de datos y utilizarla de manera efectiva para aumentar nuestro conocimiento, orientar la toma de decisiones y dirigir la mejora de procesos y servicios.

Esta breve revisión de la evolución de la estadística muestra cómo los grandes períodos de avances teóricos se han producido generalmente ligados a la resolución de importantes problemas prácticos. Paradójicamente, la mayoría de las contribuciones importantes que hemos revisado son debidas

a investigadores que no pueden calificarse exclusivamente como estadísticos, sino, en el sentido más amplio del término, como científicos. Además, los avances más importantes en algunas disciplinas se han producido por la utilización de métodos generados para resolver los problemas en otra. Por ejemplo, métodos estadísticos desarrollados para estudiar los procesos de difusión en física están revolucionando la investigación financiera (sus contribuciones han sido reconocidas ya con un premio Nobel); herramientas desarrolladas en ingeniería espacial (el filtro de Kalman) son ya de uso común en el control automático de procesos industriales, en la predicción de magnitudes macroeconómicas y en el tratamiento digitalizado de imágenes tomadas por un escáner en medicina; y métodos de clasificación desarrollados en antropología son la base de los sistemas de concesión automática de créditos, de las máquinas que reconocen billetes y monedas, de los sistemas de reconocimiento de voz y de la construcción de buscadores eficaces en Internet.

1.5 Lecturas recomendadas

Un libro excelente sobre cómo prevenir la manipulación con la estadística es Huff (1993). Bartholomew y Bassett (1971) y Moroney (1990) presentan introducciones no técnicas a la utilización de la estadística en el mundo de hoy. Tanur *et al.* (2007) es una colección muy interesante de aplicaciones. Borel (1998) es un breve y delicioso ensayo sobre las probabilidades en la vida diaria. Huff (1993) y Kitaigorodski (1976) presentan numerosos ejemplos en la misma dirección. Rao (2004) es un interesante ensayo sobre el papel de la estadística en la investigación científica.

El lector interesado en la historia del cálculo de probabilidades puede acudir a David (1998) y Todhunter (2007). La historia de la estadística se presenta en Pearson y Kendall (1976), Kendall y Plackett (1977), Stigler (1990, 2002) y Hald (1990, 1998). Sánchez-Lafuente (1975) estudia la historia de la estadística en España hasta 1900. Biografías de estadísticos célebres de especial interés son Box (1978), que describe la vida de Fisher, Reid (1982), que narra la de Neyman, o Pearson (1990), que se centra en la de Student, y Sánchez y Valdés (2003) la de Kolmogorov. Gani (1982) contiene ensayos autobiográficos de estadísticos actuales. El libro de Bernstein (1998) es una historia fascinante del riesgo y la incertidumbre.

Primera parte

Datos

2. La descripción de una variable



Pierre-Simon, marqués de Laplace (1749-1827)
Científico francés y uno de los creadores de la teoría de la probabilidad. Hizo también contribuciones fundamentales a la física, la astronomía y las matemáticas. Fue profesor de Napoleón en la escuela militar, y aquél le eligió como ministro del Interior. Luis XVIII le hizo marqués en la restauración monárquica.

2.1 Datos y distribuciones de frecuencias

Dado un conjunto de datos de una variable x, la estadística descriptiva estudia procedimientos para sintetizar la información que contienen.

Los tipos de variables que consideraremos son:

a) *Variables cualitativas, categóricas o atributos:* no toman valores numéricos y describen cualidades. Por ejemplo, clasificar personas por el color de su pelo.
b) *Variables cuantitativas discretas:* toman únicamente valores enteros; corresponden en general a contar el número de veces que ocurre un suceso. Por ejemplo, número de compras de un producto en un mes.
c) *Variables cuantitativas continuas:* toman valores en un intervalo; corresponden a medir magnitudes continuas. Por ejemplo, tiempo entre la llegada de dos autobuses.

Supondremos que el orden en que se recogen los datos es irrelevante. Cuando los datos se observan con una pauta temporal fija (cada mes, año, etc.), constituyen una serie temporal y su análisis requiere métodos especiales que tengan en cuenta que el orden de los datos es informativo.

La presentación de un conjunto de datos suele hacerse indicando los valores de la variable y sus frecuencias de aparición, tanto absolutas como relativas. La *frecuencia relativa* de un suceso A se define por:

$$fr(A) = \frac{\text{número de veces que se observa } A}{\text{número total de datos}}$$

2.1.1 Distribuciones de frecuencias

La tabla 2.1 presenta un ejemplo de una distribución de frecuencias para una variable cualitativa: se indican las clases o atributos y sus frecuencias observadas. Cuando los atributos no corresponden a una escala ordinal (por ejemplo alto, medio, bajo), conviene ordenarlos por su frecuencia de aparición.

Tabla 2.1 Distribución de defectos en libros en una imprenta

Clases	Frecuencia	Frecuencia relativa
Corte de las hojas	60	0,43
Mala impresión	40	0,29
Tinta irregular	20	0,14
Encuadernación	12	0,09
Portada	6	0,04
Lomo	2	0,01
TOTAL	140	1

La tabla 2.2 presenta esta misma idea para una variable discreta. Esta representación es útil cuando el número de valores posibles es pequeño. En otro caso, conviene agrupar los datos, como se aprecia en la mencionada tabla.

Agrupamiento

Cuando el número de valores distintos que toma una variable discreta sea grande, o cuando ésta sea continua, conviene agrupar los datos en clases, como sigue:

Tabla 2.2 Distribución de frecuencias de la variable: número de llamadas recibidas en una centralita en períodos de un minuto

X	(*f*) frecuencia	(*fr*) frecuencia relativa
0	40	0,44
1	26	0,29
2	14	0,16
3	6	0,07
4	3	0,03
5	0	0,00
6	1	0,01
TOTAL	90	1

a) Redondear los datos a dos o, a lo sumo, tres cifras significativas eligiendo las unidades para que cada observación contenga dos o tres dígitos, sin coma decimal.
b) Decidir el número r de clases a considerar. Este número debe ser entre 5 y 20. Una regla frecuentemente utilizada es tomar r igual al entero más próximo a $\sqrt{n}$, siendo n el número de datos. Esta regla es indicativa y conviene probar con distinto número de clases y escoger aquel que proporcione una descripción más clara.
c) Seleccionar los límites de clase que definen los intervalos, de manera que las clases sean de la misma longitud y cada observación se clasifique sin ambigüedad en una sola clase.
d) Contar el número de observaciones en cada clase, que llamaremos la frecuencia de clase, y obtener la frecuencia relativa de cada clase dividiendo aquélla por el total de datos.

La tabla 2.3 presenta un ejemplo de una distribución de frecuencias para una variable continua. Llamaremos en adelante *marca de clase* al centro del intervalo que define la clase.

2.1.2 Diagramas de tallo y hojas

Un procedimiento semigráfico de presentar la información para variables cuantitativas, que es especialmente útil cuando el número total de datos es pequeño (menor que 50), es el diagrama de tallo y hojas de Tukey. Los principios para construirlo son:

Tabla 2.3 Distribución de la variable: tiempo en minutos al realizar una operación

Intervalo	Centro del intervalo	Frecuencia relativa
20-24	22	0,30
25-29	27	0,40
30-34	32	0,20
35-39	37	0,07
40-44	42	0,03

a) Redondear los datos a dos o tres cifras significativas, expresándolos en unidades convenientes.
b) Disponerlos en una tabla con dos columnas separadas por una línea como sigue:

   b.1) Para datos con dos dígitos, escribir a la izquierda de la línea los dígitos de las decenas —que forma el tallo— y a la derecha las unidades, que serán las hojas. Por ejemplo, 87 se escribe 8|7.
   b.2) Para datos con tres dígitos el tallo estará formado por los dígitos de las centenas y decenas, que se escribirán a la izquierda, separados de las unidades. Por ejemplo, 127 será 12|7.

c) Cada tallo define una clase, y se escribe sólo una vez. El número de «hojas» representa la frecuencia de dicha clase.

La tabla 2.4 presenta un ejemplo de estos diagramas.

Cuando el primer dígito de la clasificación varía poco, la mayoría de los datos tienden a agruparse alrededor de un tallo y el diagrama resultante tiene poco detalle. En ese caso es conveniente subdividir cada tallo en dos o más partes introduciendo algún signo arbitrario, como se indica en la tabla 2.5.

2.2 Representaciones gráficas

2.2.1 Diagrama de Pareto

Este diagrama se utiliza para representar datos cualitativos y se construye como sigue: