

¡El libro sobre estadística para todos!

Estadística

PARA

DUMMIES[®]

Aprende a:

- Entender los cálculos, fórmulas y conclusiones estadísticas
- Analizar gráficos y tablas, determinar las probabilidades y detectar las interpretaciones sesgadas
- Comprender los datos de encuestas, sondeos, investigaciones y demás estudios sociológicos

Deborah J. Rumsey

*Profesora de Estadística en la
Universidad Estatal de Ohio
(Estados Unidos)*



Estadística

PARA
DUMMIES™

Debora J. Rumsey

Traducción de Alfredo García Espada

Sumario

Introducción	1
Acerca de este libro	1
Convenciones utilizadas en este libro	2
Lo que no necesitas leer	3
Algunas presuposiciones	3
Cómo está organizado el libro	3
Parte I: Estadísticas básicas sobre la estadística	4
Parte II: Nociones sobre cálculo numérico	4
Parte III: Distribuciones y el teorema del límite central	4
Parte IV: Estimaciones aproximadas y formulación de hipótesis fiables	4
Parte V: Estudios estadísticos y la búsqueda de una relación significativa	5
Parte VI: Los decálogos	6
Iconos utilizados en este libro	6
Lo que tienes por delante	7

Parte I: Estadísticas básicas sobre la estadística **9**

Capítulo 1: La estadística en pocas palabras	11
Prosperar en un mundo estadístico	12
Diseñar estudios adecuados	13
Encuestas	13
Experimentos	14
Recopilar datos de calidad	15
Seleccionar una muestra válida	15
Evitar el sesgo en los datos	16
Crear resúmenes eficaces	17
Estadísticos descriptivos	17
Diagramas y gráficos	18
Determinar distribuciones	19
Realizar análisis adecuados	20
Margen de error e intervalos de confianza	20
Contrastes de hipótesis	22
Correlación, regresión y tablas de contingencia	23
Extraer conclusiones fiables	24

Detectar resultados exagerados 24
Cuestionar afirmaciones de causa-efecto..... 25
Conviértete en un detective, no en un escéptico 25

Capítulo 2: La estadística de la vida cotidiana 27

Estadísticas en los medios de comunicación: ¿más preguntas que respuestas? 28
 Problemas con las palomitas 28
 La amenaza vírica 29
 Entender los accidentes 29
 A vueltas con las negligencias médicas 30
 El problema de perder terreno 31
 Los colegios en el punto de mira 31
 Los números del deporte 32
 Información económica 33
 El viaje de las cifras 34
 Las estadísticas del sexo 34
 Los misterios del parte meteorológico 35
 Reflexiones sobre las películas 35
 Los horóscopos, en el punto de mira 36
Estadísticas en el trabajo 36
 Recién nacidos e información 37
 Posar o no posar 37
 Datos sobre pizzas 38
 Estadísticas en la oficina 38

Capítulo 3: Tomar el control: tantos números y tan poco tiempo 39

Detectar errores, exageraciones y mentiras 40
 Comprobar las operaciones 40
 Descubrir estadísticas engañosas 41
 Buscar las mentiras en los lugares adecuados 48
El impacto de las estadísticas engañosas 49

Capítulo 4: Herramientas del oficio 51

Estadística: algo más que números 51
Terminología básica de estadística 53
 Datos 54
 Conjunto de datos 55
 Variable 55
 Población 56
 Muestras y aleatoriedad 57
 Estadístico 59
 Parámetro 59

Sesgo	60
Media (promedio)	61
Mediana	61
Desviación estándar (o típica)	61
Percentil	62
Puntuación estándar (o típica)	63
Distribución y distribución normal	63
Teorema del límite central	65
Valores z	65
Experimentos	66
Encuestas (sondeos)	68
Margen de error	68
Intervalo de confianza	69
Contrastes de hipótesis	70
Valores p	71
Significación estadística	72
Correlación y causalidad	73

Parte II: Nociones de cálculo numérico 75

Capítulo 5: Medias, medianas y más 77

Resumir datos con estadísticos descriptivos	77
Cálculos con datos categóricos: tablas y porcentajes	78
Medir el centro con la media y la mediana	81
Calcular la media	81
Calcular la mediana	83
Comparar medias y medianas: histogramas	85
Tener en cuenta la variación	87
Comunicar la desviación estándar	87
Fuera de rango	91
La regla empírica (68-95-99,7)	92
Medir la posición relativa con percentiles	95
Calcular percentiles	95
Interpretar percentiles	96
Hacer un resumen pentanumérico	101
Consideraciones sobre el rango intercuartílico	102

Capítulo 6: Representación gráfica de datos categóricos 103

Coge otro trocito de tarta	104
Contabilidad de gastos personales	104
Los ingresos del juego	104
Comida para llevar	106
Proyección de tendencias de edad	107

Barra libre de gráficos de barras	109
Gasto en transporte	109
Los beneficios del juego	111
Escalas y gráficos de barras	112
Fastidios cotidianos	113

Capítulo 7: Representación gráfica de datos numéricos 117

Historia de un histograma	117
Elaborar un histograma	118
Interpretar un histograma	121
Relacionar números y gráficos	125
Detectar histogramas engañosos	127
Diagramas de cajas	131
Elaborar un diagrama de cajas	131
Interpretar un diagrama de cajas	132
Tiempo de cronogramas	138
Interpretar cronogramas	139
Comprender la variabilidad: cronogramas e histogramas	140
Detectar cronogramas engañosos	140

Parte III: Distribuciones y el teorema del límite central .. 145

Capítulo 8: Variables aleatorias y la distribución binomial 147

Definir una variable aleatoria	147
Discreto y continuo	148
Distribuciones de probabilidad	149
La media y la variancia de una variable aleatoria discreta	150
Identificar una variable binomial	151
Comprobar paso a paso las condiciones de distribución binomial	152
El número de intentos no es fijo	153
Hay más resultados, aparte de éxito o fracaso	153
Los intentos no son independientes	153
La probabilidad de éxito (p) varía	154
Encontrar probabilidades binomiales utilizando una fórmula	154
Encontrar probabilidades con la tabla binomial	157
Encontrar probabilidades para valores concretos de X	157
Encontrar probabilidades para X mayor que un valor, menor que un valor o entre dos valores	158
Comprobar la media y la desviación estándar de la distribución binomial	159

Capítulo 9: La distribución normal 161

Fundamentos de la distribución normal	161
La distribución normal estándar (distribución Z)	164

Particularidades de la distribución Z	165
Normalizar de X a Z	166
Encontrar probabilidades para Z con la tabla Z	167
Encontrar probabilidades para una distribución normal	168
Encontrar X cuando conoces el porcentaje	171
Determinar un percentil para una distribución normal	172
Un problema de percentil bajo	172
Problemas de percentiles difíciles de interpretar	174
Aproximación normal a la distribución binomial	175
Capítulo 10: La distribución t	177
Aspectos básicos de la distribución t	177
Comparación entre la distribución t y la distribución Z	177
Efecto de la variabilidad en las distribuciones t	179
Uso de la tabla t	180
Encontrar probabilidades con la tabla t	180
Encontrar percentiles para la distribución t	181
Escoger valores t^* para intervalos de confianza	181
Estudiar comportamientos con la tabla t	182
Capítulo 11: Distribuciones muestrales y el teorema del límite central 185	
Definir una distribución muestral	185
La media de una distribución muestral	186
Medir el error estándar	188
Tamaño muestral y error estándar	189
Desviación estándar de la población y error estándar	190
Examinar la forma de una distribución muestral	192
Caso 1: la distribución de X es normal	192
Caso 2: la distribución de X no es normal (entra en escena el teorema del límite central)	193
Encontrar probabilidades para la media muestral	196
La distribución muestral de la proporción muestral	198
Encontrar probabilidades para la proporción muestral	201
<i>Parte IV: Estimaciones aproximadas y formulación de hipótesis fiables</i>	203
Capítulo 12: Prever un margen de error	205
La importancia del más/menos	205
Fórmula general para encontrar el margen de error	207
Medir la variabilidad de la muestra	207
Calcular el margen de error para una proporción muestral	209
Comunicar resultados	211
Calcular el margen de error para una media muestral	212
Cuestión de confianza	214

Repercusiones del tamaño muestral	214
Tamaño muestral y margen de error	214
No por ser más grande es (mucho) mejor	215
No perder la perspectiva del margen de error	216

Capítulo 13: Intervalos de confianza: estimaciones aproximadas 219

No todas las estimaciones son iguales	219
Vincular un estadístico a un parámetro	221
Terminología básica	222
Interpretar resultados con confianza	222
Cuestión de amplitud	223
Elegir un nivel de confianza	225
La importancia del tamaño muestral	226
Tener en cuenta la variabilidad de la población	228
Calcular un intervalo de confianza para una media poblacional	229
Caso 1: Se conoce la desviación estándar de la población	229
Caso 2: Se desconoce la desviación estándar de la población y/o n es pequeño	230
Determinar qué tamaño muestral necesitas	232
Determinar el intervalo de confianza para una proporción poblacional	233
Crear un intervalo de confianza para la diferencia entre dos medias	235
Caso 1: Se conocen las desviaciones estándares de la población	236
Caso 2: Se desconocen las desviaciones estándares de la población y/o los tamaños muestrales son pequeños.....	237
Estimar la diferencia entre dos proporciones	239
Detectar intervalos de confianza engañosos	241

Capítulo 14: Afirmaciones, contrastes y conclusiones 243

Formular las hipótesis	244
Definir la hipótesis nula	245
¿Cuál es la alternativa?	245
Recopilar pruebas (datos)	246
Procesar las pruebas: el estadístico de contraste	247
Recopilar estadísticos muestrales	247
Medir la variabilidad con errores estándares	247
Comprender las puntuaciones estándares	248
Calcular e interpretar el estadístico de contraste	248
Ponderar los indicios y tomar decisiones: valores p	249
Relacionar estadísticos de contraste y valores p	250
Definir un valor p	251
Calcular un valor p	251
Extraer conclusiones	252
Establecer límites para rechazar H_0	254

Un problema de varices 254
 Valorar la posibilidad de tomar una decisión errónea 255
 Provocar una falsa alarma: errores de tipo 1 255
 No detectar una afirmación falsa: errores de tipo 2 256

**Capítulo 15: Contrastes de hipótesis frecuentemente utilizados:
 fórmulas y ejemplos 259**

Contrastar una media poblacional 260
 Muestras pequeñas y desviaciones estándares desconocidas: la
 prueba *t* 262
 Aplicación práctica de la prueba *t* 263
 Relacionar *t* y *Z* 264
 Qué hacer con valores *t* negativos 264
 Analizar la alternativa “distinto de” 265
 Contrastar una proporción poblacional 265
 Comparar dos medias poblacionales (independientes) 267
 Contraste de hipótesis para una diferencia media (prueba *t* para
 datos apareados) 270
 Comparar dos proporciones poblacionales 274

***Parte V: Estudios estadísticos y la búsqueda de una
 relación significativa 277***

Capítulo 16: Encuestas, encuestas y más encuestas 279

Comprender la importancia de las encuestas 280
 Llegar hasta la fuente 280
 Preguntar sobre temas candentes 282
 Repercusión sobre las personas 283
 Entre bastidores: los entresijos de las encuestas 284
 Planificar y diseñar una encuesta 285
 Seleccionar la muestra 289
 Realizar la encuesta 292
 Interpretar los resultados y descubrir problemas 295

**Capítulo 17: Experimentos: ¿logros médicos extraordinarios
 o resultados engañosos? 299**

Consideraciones básicas sobre los estudios 300
 Terminología técnica 300
 Observar los estudios de observación 301
 Examinar los experimentos 302
 Diseñar un experimento válido 302
 Diseñar el experimento para realizar comparaciones 303

- Seleccionar el tamaño muestral 306
- Seleccionar a los sujetos 307
- Realizar asignaciones aleatorias 308
- Controlar las variables de confusión 309
- Problemas éticos 311
- Recopilar datos válidos 312
- Analizar los datos correctamente 314
- Extraer conclusiones apropiadas 315
- Tomar decisiones con conocimiento de causa 317

Capítulo 18: Buscar vínculos: correlación y regresión 319

- Representar una relación con un diagrama de dispersión 320
 - Elaborar un diagrama de dispersión 321
 - Interpretar un diagrama de dispersión 322
- Cuantificar relaciones lineales por medio de la correlación 323
 - Calcular la correlación 323
 - Interpretar la correlación 325
 - Examinar las propiedades de la correlación 327
- Trabajar con la regresión lineal 327
 - Determinar qué variable es X y cuál es Y 328
 - Comprobar las condiciones 328
 - Calcular la recta de regresión 329
 - Interpretar la recta de regresión 331
 - Un ejemplo ilustrativo: la recta de regresión para los grillos ... 333
- Realizar predicciones adecuadas 333
- Explicar la relación: correlación frente a causa y efecto 335

Capítulo 19: Tablas de contingencia e independencia 337

- Organizar una tabla de contingencia 338
 - Crear las celdas 339
 - Calcular los totales 340
- Interpretar las tablas de contingencia 341
 - Análisis de variables aisladas con distribuciones marginales .. 342
 - Analizar todos los grupos: una distribución conjunta 345
 - Comparar grupos con distribuciones condicionales 348
- Comprobar la independencia y describir la dependencia 352
 - Comprobar la independencia 352
 - Describir una relación de dependencia 356
- Interpretar los resultados con precaución 357
 - Comprobar si existe una relación de causa y efecto válida 358
 - Extrapolar de la muestra a la población 359
 - Realizar predicciones prudentes 360
 - Resistir el impulso de extraer conclusiones precipitadas 361

Parte VI: Los decálogos 363**Capítulo 20: Diez consejos para ser un sabueso
de las estadísticas 365**

Detectar gráficos engañosos.....	365
Gráficos de sectores.....	366
Gráficos de barras.....	366
Cronogramas.....	367
Histogramas.....	368
Descubrir datos sesgados.....	370
Buscar un margen de error.....	371
Identificar muestras no aleatorias.....	371
El tamaño sí que importa.....	372
Detectar correlaciones malinterpretadas.....	373
Revelar variables de confusión.....	374
Examinar los números.....	375
Informar de los informes selectivos.....	376
Distinguir las anécdotas.....	377

Capítulo 21: Diez consejos para sacar mejor nota en los exámenes.. 379

Averiguar qué es lo que no sabes, y ponerle remedio.....	380
Evitar las trampas «Ya me lo sé».....	381
Trampas «Ya me lo sé» n.º 1.....	382
Trampas «Ya me lo sé» n.º 2.....	383
Hacerse amigo de las fórmulas.....	384
Elaborar una tabla procedimental.....	386
Entender qué te están preguntando.....	388
Etiquetar la información que te dan.....	389
Hacer un dibujo.....	391
Establecer la conexión y resolver el problema.....	392
Haz las cuentas... dos veces.....	393
Analizar tus respuestas.....	394

Apéndice: Tablas de referencia..... 395

La tabla Z.....	395
La tabla t.....	398
La tabla binomial.....	400

Índice 405

Capítulo 1

La estadística en pocas palabras

En este capítulo

- ▶ Descubrir en qué consisten los procesos estadísticos
- ▶ Aprovechar la estadística para triunfar en la vida cotidiana, en el terreno profesional y en los estudios

Hoy en día nos bombardean con tal cantidad de datos que cualquiera (incluso yo) puede sentirse abrumado. Desde luego, yo no te culparía por desconfiar de las estadísticas que lees a diario en los medios. A mí me ocurre lo mismo. Lo bueno es que, a pesar de la gran cantidad de información engañosa e incorrecta que se publica, también se produce material de calidad; por ejemplo, muchos estudios y técnicas relacionadas con datos ayudan a mejorar la calidad de vida de las personas. Tu tarea es distinguir lo bueno de lo malo y confiar en tu capacidad de lograrlo. Al entender los pormenores de la estadística y los procedimientos estadísticos, ganas confianza con los números y cobras ventaja en tu vida cotidiana, en el trabajo y en los estudios. Y de eso justamente trata este libro.

En este capítulo te ofrezco una visión general del papel que desempeña la estadística en la sociedad actual, rebosante de datos y cifras, y te cuento lo que puedes hacer para no sólo sobrevivir, sino prosperar. Obtendrás una perspectiva mucho más amplia de la estadística si te adentras en el método científico y aprendes a diseñar estudios eficaces, recopilar datos válidos, organizar y analizar la información, interpretar los resultados y extraer conclusiones adecuadas (¡y tú que pensabas que la estadística era sólo hacer cálculos!).

Prosperar en un mundo estadístico

Todos los días recibimos un aluvión de estadísticas que cuesta mucho controlar y nos afectan de maneras muy distintas. Comienza por la mañana, nada más salir de la cama, cuando enciendes el televisor y el hombre del tiempo te ofrece unas predicciones basadas en análisis estadísticos de datos antiguos y condiciones meteorológicas actuales. Mientras desayunas, lees atentamente la información nutricional impresa en la caja de cereales. En el trabajo te enfrentas a diagramas y tablas repletas de números, introduces datos en hojas de cálculo, realizas diagnósticos, tomas medidas, estimas gastos, tomas decisiones a partir de estadísticas y pides presupuestos teniendo en cuenta los datos de venta de los meses anteriores.

A mediodía vas a comer al mejor restaurante de la zona según una encuesta realizada a 500 personas. El precio de los platos que te sirven se determina tras consultar datos de marketing. En la consulta del médico te miden la tensión arterial, la temperatura corporal y el peso, te hacen un análisis de sangre y después te dan un informe donde tus resultados aparecen comparados con los valores que se consideran normales desde un punto de vista estadístico.

Conduces un coche que han puesto a punto con un ordenador que ejecuta diagnósticos estadísticos. Cuando llegas a casa, pones las noticias y oyes las últimas estadísticas sobre delincuencia, te enteras de cómo ha ido la Bolsa y descubres cuántas personas fueron al zoo la semana pasada.

Por la noche te cepillas los dientes con un dentífrico que, según las estadísticas, es muy eficaz contra la caries. Lees unas cuantas páginas de un libro que te compraste tras ver unas estimaciones de ventas basadas en criterios estadísticos, y por fin te duermes. Y a la mañana siguiente vuelta a empezar. Pero ¿cómo puedes estar seguro de que todas esas estadísticas en las que confías a diario son correctas? En el capítulo 2 comento con más detalle algunos ejemplos de cómo la estadística está presente en nuestras vidas personales y profesionales, en qué medida te afecta y cómo puedes ser más consciente de ella.



Algunas estadísticas son vagas, inadecuadas o simplemente erróneas. Debes tomar conciencia de las cifras estadísticas que te encuentras a diario y entrenar la mente para ser capaz de distanciarte de toda esa información, examinarla con detenimiento, hacerte preguntas y pulsar el botón de alarma cuando algo no termine de cuadrar. En el capítulo 3 te doy varios ejemplos de cómo puede engañarte una estadística errónea y te enseño a desarrollar la capacidad de pensar de forma crítica e identificar problemas antes de creerte los resultados a pies juntillas.

Igual que cualquier otro campo del saber, la estadística tiene su propia jerga técnica. En el capítulo 4 verás el significado de algunos de los términos estadísticos más utilizados. Dominar la terminología te hace más capaz de comprender y transmitir estadísticas de cierta complejidad sin sentirte intimidado. Este conocimiento aumenta tu credibilidad cuando utilizas términos precisos para señalar (y explicar) un error en un resultado estadístico. Además, cuando utilices tablas, gráficos y análisis estadísticos, tus presentaciones transmitirán más información y serán más efectivas (bueno, el caso es que debes conocer toda esa jerga porque la utilizo en el libro; pero no te preocupes porque siempre te la recuerdo).

En los siguientes apartados verás de qué modo la estadística interviene en todas las fases del método científico.

Diseñar estudios adecuados

Todo el mundo hace preguntas, desde las compañías farmacéuticas hasta los biólogos, pasando por los analistas de marketing y los gobiernos. Y, en último término, todo el mundo utiliza las estadísticas para responder a esas preguntas. En concreto, muchos estudios médicos y psicológicos se hacen porque alguien quiere saber la respuesta a una determinada pregunta. Por ejemplo,

- ✓ ¿Esta vacuna será efectiva para prevenir la gripe?
- ✓ ¿Qué piensan los estadounidenses sobre la coyuntura económica actual?
- ✓ ¿El mayor uso de las redes sociales es causa de depresión entre los adolescentes?

El primer paso tras plantear la pregunta que da pie a la investigación consiste en diseñar un estudio efectivo para recopilar datos que ayuden a responder a esa pregunta. Este paso equivale a decidir qué proceso debes utilizar para obtener los datos que necesitas. En este apartado comento las principales características de los dos tipos de estudios más importantes (encuestas y experimentos) y por qué es tan importante valorar la manera en que se ha diseñado un estudio antes de creerse los resultados.

Encuestas

Un *estudio de observación* es aquel en que se recopilan datos de personas sin que éstas se vean afectadas. El estudio de observación más común es la encuesta. Las *encuestas* son cuestionarios que se presentan a personas

14 Parte I: Estadísticas básicas sobre la estadística

escogidas del grupo de población que es objeto de estudio. Las encuestas pueden adoptar distintas formas: encuestas en papel enviadas por correo, cuestionarios en sitios web, encuestas de llamada voluntaria realizadas por cadenas de televisión, encuestas telefónicas, etc.



Si se realizan correctamente, las encuestas son unas herramientas muy útiles para obtener información. Sin embargo, si no se hacen bien, pueden proporcionar datos erróneos. Algunos problemas pueden provenir de errores en la formulación de las preguntas (que pueden dar lugar a equívocos), de la negativa a responder por parte de las personas que fueron elegidas para participar, o de no incluir todos los grupos de población. Debido a estos posibles problemas, es imprescindible meditar a fondo la encuesta antes de realizarla.



Muchos investigadores invierten mucho tiempo y dinero en diseñar buenas encuestas (según los criterios que expongo en el capítulo 16), y entonces sí se puede confiar en ellas. No obstante, como te encuentras con encuestas de todo tipo en los medios de comunicación, en el trabajo y en muchas de tus clases, es importante que sepas evaluar rápidamente el modo en que se han diseñado y realizado dichas encuestas y puedas identificar problemas concretos con conocimiento de causa. Las herramientas que necesitas para valorar una encuesta se encuentran en el capítulo 16.

Experimentos

Un *experimento* impone uno o más tratamientos a los participantes de manera que puedan establecerse comparaciones claras. Una vez aplicados los tratamientos, se anotan las respuestas. Por ejemplo, para estudiar el efecto de la dosis de un fármaco sobre la tensión arterial, un grupo puede tomar 10 miligramos de dicho fármaco y otro grupo puede tomar 20 miligramos. Generalmente también hay un grupo de control, cuyos integrantes reciben un placebo (una pastilla de azúcar) o un tratamiento estándar no experimental (como los medicamentos que se administran actualmente a los enfermos de Sida).



Para ser válido y fiable, un experimento debe diseñarse de manera que se reduzca al mínimo el sesgo, que se recopilen muchos datos válidos y que se realicen comparaciones adecuadas (entre el grupo experimental y el grupo de control). Entre los posibles problemas que puede haber con un experimento están los siguientes: que los investigadores o los sujetos conozcan el tratamiento, que el resultado del estudio se vea afectado por factores no controlados (por ejemplo el peso del sujeto, si se está estudiando la dosis de un fármaco) o que no haya un grupo de control (en cuyo caso no existe una referencia con la que comparar los resultados).

Pero, si se diseña como es debido, un experimento puede ayudar al investigador a establecer una relación causa-efecto si la diferencia entre las respuestas del grupo experimental y del grupo de control es estadísticamente significativa (es decir, si es improbable que haya ocurrido de forma casual).



Los experimentos ayudan a crear y probar nuevos fármacos, establecer normas para la manipulación de alimentos y evaluar si un nuevo tratamiento médico puede curar una enfermedad o al menos reducir sus consecuencias. Sin duda, nuestra calidad de vida ha mejorado gracias al uso de experimentos bien diseñados. Sin embargo, no todos lo están, y por eso es fundamental que sepas distinguir qué resultados son creíbles y cuáles son increíbles (esto es un chiste), sobre todo cuando las conclusiones son importantes para ti. En el capítulo 17 encontrarás todo lo que debes saber sobre los experimentos y la forma de evaluarlos.

Recopilar datos de calidad

Una vez diseñado un estudio, ya sea una encuesta o un experimento, hay que seleccionar los participantes y tener claro el modo en que se van a recopilar los datos. Esta fase del proceso es clave para obtener datos fiables, y de eso trata este apartado.

Seleccionar una muestra válida



Los estadísticos tienen un dicho: “Si entra basura, sale basura”. Si seleccionas los *sujetos* (las personas que participarán en el estudio) de manera *sesgada* (es decir, favoreciendo a determinadas personas o grupos de personas), los resultados también estarán sesgados. Así de simple.

Pongamos que Miguel quiere conocer la opinión de los habitantes de tu ciudad acerca de un casino que podría abrirse. Miguel va a un centro comercial con una carpeta sujetapapeles y se dedica a preguntar a la gente que pasa por su lado. ¿Está haciendo algo mal? Bueno, sólo va a conseguir las opiniones de a) las personas que compren en ese centro comercial; b) ese día en concreto; c) a esa hora en concreto; d) y que se tomen la molestia de contestar.

Esas circunstancias son demasiado restrictivas, ya que los sujetos a los que pregunta no representan a una sección transversal de la ciudad. Miguel también podría colgar la encuesta en Internet y pedir a la gente que entre en la web para votar. No obstante, tan sólo obtendría datos de quienes conocieran la web, tuvieran acceso a Internet y además quisieran respon-

16 Parte I: Estadísticas básicas sobre la estadística

der, y suele ocurrir que sólo las personas con opiniones muy tajantes se toman tantas molestias. Al final Miguel no tiene más que un montón de datos sesgados sobre personas que no representan en absoluto a la ciudad.



Para reducir al mínimo el sesgo en una encuesta, la palabra clave es *aleatoriedad*. La muestra de personas debe elegirse de forma *aleatoria*, es decir, al azar, como si se sacaran los nombres de un sombrero. Los científicos utilizan varios métodos para elegir personas de forma aleatoria (los encontrarás en el capítulo 16).

Ten en cuenta que, a la hora de diseñar un experimento, generalmente no es ético coger una muestra aleatoria de personas y pedirles que participen, ya que los experimentos obligan a los sujetos a seguir un tratamiento. Lo que se hace entonces es pedir voluntarios. Hay que asegurarse de que los voluntarios escogidos representen a la población de interés y que sus datos se recopilen correctamente para luego poder extrapolar los resultados a un grupo más grande. El procedimiento se explica en el capítulo 17.

Después de leer los capítulos 16 y 17, habrás aprendido a analizar los métodos de otros para selección de muestras e incluso sabrás diseñar un plan propio. Al final sabrás cuándo decir: "Si entra basura, sale basura".

Evitar el sesgo en los datos

El *sesgo* es el favoritismo sistemático hacia determinadas personas o determinadas respuestas. El sesgo es el principal enemigo de los estadísticos, que hacen todo lo posible por reducirlo al mínimo. ¿Quieres un ejemplo de sesgo? Imagina que estás haciendo una encuesta telefónica sobre satisfacción en el trabajo. Si haces las llamadas a los domicilios entre las nueve de la mañana y las cinco de la tarde, no hablarás con nadie que trabaje durante el día. Puede que los trabajadores diurnos estén más satisfechos que los nocturnos.

Cuando se recopilan datos para una encuesta, hay que evitar el sesgo. Por ejemplo: algunas encuestas son demasiado largas, ¿qué pasa si alguien deja de responder a las preguntas a la mitad del cuestionario? ¿Y si el encuestado miente y te dice que gana 100.000 euros en lugar de 45.000? ¿Y si te da una respuesta que no está en la lista de respuestas posibles? Cuando se están recopilando datos en una encuesta pueden surgir muchos problemas, y tú tienes que ser capaz de detectarlos.



A veces los experimentos plantean incluso más problemas que las encuestas en lo que respecta al sesgo y la recopilación de datos. Pongamos que quieres medir la tensión arterial. ¿Qué pasa si el instrumento que utilizas se rompe en mitad del experimento? ¿Y si alguien abandona el experimen-

to a medias? ¿Y si durante el experimento ocurre algo que distrae a los sujetos o a los investigadores? ¿Y si tienen que hacer un análisis de sangre exactamente una hora después de haber administrado una dosis de un fármaco y no logran encontrar una vena? Éstos son sólo algunos ejemplos de cosas que pueden salir mal cuando estás recopilando datos para un experimento, y debes estar preparado para localizar estos problemas.

Cuando hayas terminado de leer el capítulo 16 (sobre muestras y encuestas) y el capítulo 17 (sobre experimentos), sabrás seleccionar muestras y recopilar datos de manera no sesgada, prestando atención a todas esas pequeñas cosas que pueden influir en los resultados. Además, habrás aprendido a valorar la fiabilidad de resultados estadísticos y podrás hacerle oír porque sabrás de qué estás hablando.

Crear resúmenes eficaces

Una vez que has recopilado datos válidos, el siguiente paso consiste en resumirlos para tener una visión más amplia. Los estadísticos describen los datos de dos formas distintas: mediante números (llamados *estadísticos descriptivos*) y mediante dibujos (es decir, diagramas y gráficos).

Estadísticos descriptivos

Los *estadísticos descriptivos* son números que describen las características importantes de un conjunto de datos:



- ✓ Si los datos son *categoricos* (cuando las personas se enmarcan en grupos, por ejemplo por razón de su sexo o filiación política), generalmente se resumen utilizando el número de personas que hay en cada grupo (la *frecuencia*) o el porcentaje de personas en cada grupo (la *frecuencia relativa*).
- ✓ Los *datos numéricos* representan medidas o recuentos, de modo que los números tienen significado por sí mismos (por ejemplo la altura y el peso). Los datos numéricos permiten resumir otras características, aparte del número o porcentaje que hay en cada grupo. Estas características incluyen, entre otras:
 - Medidas de tendencia central (dicho de otro modo, ¿dónde está el “centro” de los datos?).
 - Medidas de dispersión (¿qué grado de diversidad o concentración presentan los datos en torno al centro?).
 - Si procede, números que midan la relación entre dos variables (por ejemplo, la altura y el peso).



Algunos estadísticos descriptivos resultan más adecuados que otros en determinadas situaciones. Por ejemplo, la media no es siempre la mejor medida del centro de un conjunto de datos, sino que la mediana es a menudo una elección más acertada. Asimismo, la desviación estándar no es la única medida de variabilidad que existe: el rango intercuartílico también resulta muy adecuado. Tienes que aprender a discriminar, interpretar y valorar los tipos de estadísticos descriptivos que ves a diario, y saber cuándo convendría utilizar un estadístico más apropiado.

Los estadísticos descriptivos más comunes se calculan, interpretan, comparan y valoran en el capítulo 5. Estos estadísticos descriptivos de uso habitual son las frecuencias y frecuencias relativas (recuentos y porcentajes), para los datos categóricos, y la media, mediana, desviación estándar, percentiles y sus combinaciones, para los datos numéricos.

Diagramas y gráficos

Para resumir los datos de forma visual se utilizan diagramas y gráficos. Se trata de presentaciones organizadas para que puedas obtener rápidamente una visión de conjunto de los datos o centrarte en un resultado en concreto. En un mundo donde prima la información rápida y las citas jugosas, los gráficos y diagramas son cosa común y corriente. La mayoría de ellos muestran la información de manera clara, eficaz e imparcial. Lo malo es que dejan cierto margen para licencias poéticas y, en consecuencia, pueden exponerte a gran cantidad de datos engañosos e incorrectos.



En los capítulos 6 y 7 hablo sobre los principales tipos de diagramas y gráficos utilizados para resumir datos categóricos y numéricos (en el apartado anterior encontrarás más información sobre ambos tipos de datos). Te explicaré cómo se elaboran, para qué sirven y cómo se interpretan los resultados. También te mostraré muchas formas de elaborar diagramas y gráficos con el propósito de engañar, y te diré cómo detectar rápidamente los problemas. Es cuestión de que aprendas a decir: “¡Espera un momento! ¡Esto no está bien!”, y sepas por qué. Te adelanto algunos puntos destacados:

- ✓ Entre los gráficos básicos utilizados para datos categóricos se encuentran los gráficos de sectores y los gráficos de barras, que desglosan variables como el sexo de la persona o las aplicaciones de móvil utilizadas por los adolescentes. Un gráfico de barras, por ejemplo, puede mostrar las opiniones sobre una cuestión utilizando cinco barras ordenadas desde “totalmente en contra” hasta “totalmente a favor”. En el capítulo 6 encontrarás abundante información sobre la manera de elaborar e interpretar y, lo que es aún más importante,

evaluar estos diagramas y gráficos para saber si son imparciales. Te sorprenderás al ver cuántos errores pueden llegar a cometerse en un simple gráfico de barras.

- ✓ Para datos numéricos como altura, peso, tiempo o cantidad se necesita un tipo de gráfico distinto. Los histogramas y los diagramas de cajas pueden ser muy explicativos y aportar información de calidad sobre un conjunto de datos. Pero también pueden llevar a equívocos, ya sea por accidente o de manera intencionada (más información en el capítulo 7).



Los diagramas y los gráficos son el pan nuestro de cada día; no tienes más que abrir un periódico para encontrarte con alguno sin tan siquiera buscar. Tener una lupa de estadístico que te ayude a interpretar la información es fundamental para detectar gráficos engañosos antes de extraer conclusiones incorrectas que puedan llevarte a tomar decisiones incorrectas. Todas las herramientas que necesitas están esperándote en el capítulo 6 (para datos categóricos) y en el capítulo 7 (para datos numéricos).

Determinar distribuciones

Una *variable* es una característica que se cuenta, se mide o se categoriza, como por ejemplo el sexo, la edad, la altura, el peso o el número de mascotas que tienes. Una *distribución* es una relación de los posibles valores (o intervalos de valores) de una variable con indicación de la frecuencia (o densidad) con que se presentan. Por ejemplo, se estima que la distribución del sexo en el momento de nacer es el 52,4% de varones y el 47,6% de mujeres.



Existen diferentes tipos de distribuciones para diferentes variables. Las tres distribuciones siguientes son las más utilizadas en un curso de introducción a la estadística, y tienen muchas aplicaciones en el mundo real:

- ✓ Si una variable cuenta los éxitos en un determinado número de intentos (por ejemplo la cantidad de personas que se curaron al tomar un determinado fármaco), tiene una distribución *binomial*.
- ✓ Si la variable adopta valores que siguen una curva en forma de campana, por ejemplo los resultados de la evaluación nacional del rendimiento académico, tiene una distribución *normal*.
- ✓ Si la variable está basada en medias muestrales y los datos son limitados, por ejemplo si se lleva a cabo un experimento con tan sólo diez sujetos para comprobar si un plan de adelgazamiento surte efecto, puedes utilizar la distribución *t*.

Tienes que saber cómo determinar la distribución que tiene una variable en particular, cómo encontrar las probabilidades y cómo calcular la desviación media y estándar de los resultados a largo plazo. Para enseñarte todo eso he incluido en el libro tres capítulos, uno dedicado a cada una de las distribuciones: el capítulo 8 trata sobre la distribución binomial, el capítulo 9 se refiere a la distribución normal y el capítulo 10 se centra en la distribución t .



Si estás haciendo un curso de introducción a la estadística (o cualquier otro curso de estadística, a decir verdad) ya sabrás que uno de los conceptos más difíciles de entender son las distribuciones muestrales y el teorema del límite central (ambas cosas van de la mano). El capítulo 11 te guía paso a paso por estos temas para que acabes entendiendo qué es una distribución muestral, qué utilidad tiene y por qué es la base para realizar análisis de datos como los contrastes de hipótesis y los intervalos de confianza (en el siguiente apartado profundizo en el análisis de datos). Cuando entiendas el teorema del límite central te será más fácil resolver problemas difíciles, y todas las claves sobre esta cuestión se encuentran en el capítulo 11.

Realizar análisis adecuados

Después de recopilar los datos y describirlos por medio de números y gráficos viene la parte divertida: adentrarnos en ese terreno pantanoso que llamamos *análisis estadístico*. Si el estudio está bien diseñado, las preguntas originales pueden responderse utilizando el análisis adecuado (la palabra clave es *adecuado*).



Existen muchos tipos de análisis y es fundamental elegir el adecuado para cada situación, así como interpretar correctamente los resultados, ser consciente de las limitaciones y saber valorar los análisis de otras personas y las conclusiones extraídas con ellos.

Este libro contiene toda la información y las herramientas que necesitas para analizar datos utilizando los métodos más habituales de la estadística básica: intervalos de confianza, contrastes de hipótesis, correlación y regresión, y análisis de tablas de contingencia. En este apartado encontrarás una visión general de todos esos métodos.

Margen de error e intervalos de confianza

Muchas veces una estadística intenta estimar números relativos a una población entera; de hecho, ves datos de ese tipo casi a diario en forma

de resultados de encuestas. Los medios de comunicación dicen cuál es el precio medio de la gasolina, qué piensan los ciudadanos sobre la actuación del gobierno o cuántas horas a la semana pasa la gente navegando por Internet.

Pero lo cierto es que nadie puede darte una cifra concreta y afirmar que es una estimación precisa de toda la población, a menos que haya recopilado datos de todas y cada una de las personas que componen esa población. Por ejemplo, pueden decirte que el 60 % de los ciudadanos no están de acuerdo con la política del gobierno en materia de sanidad, pero tú sabes que a ti no te han preguntado, de modo que ¿cómo se supone que han preguntado a todo el mundo? Y como no han preguntado a todo el mundo, sabes que dar una única cifra como respuesta no es suficiente.

Lo que en realidad ocurre es que se recopilan datos de una muestra de la población (por ejemplo, la organización Gallup llama por teléfono a 2.500 personas elegidas al azar), se analizan los resultados de esa muestra y se extraen conclusiones para toda la población (por ejemplo, todos los ciudadanos españoles).



Lo importante aquí es que los resultados son distintos para cada muestra, y esa variabilidad debe mencionarse (aunque muchas veces no se menciona). El estadístico utilizado para medir y comunicar el grado de precisión de los resultados de una muestra se llama *margen de error*. En este contexto, la palabra *error* no significa que alguien se haya equivocado; sólo significa que, al tomarse una muestra más pequeña que la población completa, existe cierta divergencia entre los resultados obtenidos y el valor real que se está intentando estimar para la población.

Por ejemplo, alguien comprueba que el 60 % de las 1.200 personas encuestadas no están de acuerdo con la política del gobierno en materia de sanidad y comunica los resultados con un margen de error de más/menos un 2%. Este intervalo de valores probables comprendido entre el 58 y el 62 % se llama *intervalo de confianza*.



Todos estamos expuestos a resultados que incluyen un margen de error e intervalos de confianza, y con la explosión de datos de hoy en día muchas personas los utilizan también en el trabajo. Es importante que conozcas los factores que influyen en el margen de error (por ejemplo el tamaño muestral) y los rasgos característicos de un buen intervalo de confianza, y también deberías ser capaz de encontrar por ti mismo un intervalo de confianza cuando lo necesites.

En el capítulo 12 encontrarás todo lo que necesitas saber sobre el margen de error: cuáles son sus componentes, qué es lo que mide y lo que no mide, y cómo se calcula para varias situaciones. El capítulo 13 te guía

paso a paso por las fórmulas, cálculos e interpretaciones de intervalos de confianza para una media poblacional, una proporción poblacional, y la diferencia entre dos medias y proporciones.

Contrastes de hipótesis

Un elemento esencial de los estudios de investigación es el contraste de hipótesis. Un *contraste de hipótesis* es una técnica que utiliza datos con el propósito de validar o invalidar una afirmación sobre una población. Por ejemplo, un político declara que el 80% de las personas de su comunidad están de acuerdo con él. ¿Está diciendo la verdad? O pongamos el caso de una empresa que asegura entregar las pizzas en treinta minutos o menos. ¿Seguro que es cierto? Los investigadores médicos utilizan los contrastes de hipótesis continuamente para comprobar si un determinado fármaco es o no eficaz, para comparar los efectos secundarios de un nuevo fármaco con los de otro fármaco ya existente, o para determinar qué plan de adelgazamiento es más eficaz para un determinado grupo de personas.



Los elementos de una población que se verifican más a menudo son:

- ✓ La media de la población (¿el tiempo medio de entrega de treinta minutos es realmente cierto?).
- ✓ La proporción de la población (¿es verdad que el 80% de los votantes apoyan a ese candidato, o la cifra es menor?).
- ✓ La diferencia entre dos medias o proporciones poblacionales (¿es cierto que el promedio de peso perdido con esta nueva dieta supera en 5 kilos al promedio de la dieta más popular? O también: ¿es cierto que este fármaco reduce la tensión arterial el 10% más que el fármaco actual?).



Los contrastes de hipótesis se utilizan en muchos ámbitos distintos que influyen en tu vida cotidiana, por ejemplo estudios médicos, anuncios, datos de sondeos y prácticamente cualquier otra aplicación donde se establezcan comparaciones tomando como base promedios o proporciones. En el terreno laboral, los contrastes de hipótesis se utilizan mucho en áreas como el marketing, para determinar si un determinado tipo de anuncio es efectivo o si un determinado grupo de personas compra más o menos tu producto en comparación con el año anterior.

Muchas veces sólo te enteras de las conclusiones de los contrastes de hipótesis (por ejemplo, este fármaco es bastante más eficaz y tiene menos efectos secundarios que el fármaco que estás utilizando ahora), pero no te dicen qué métodos han utilizado para llegar a esas conclusiones. En el capítulo 14 te cuento todos los pormenores y entresijos de los contrastes

de hipótesis, para que puedas realizarlos y valorarlos con conocimiento de causa. El capítulo 15 contiene instrucciones paso a paso para diseñar y llevar a cabo contrastes de hipótesis para un gran número de situaciones específicas (una media poblacional, una proporción poblacional, la diferencia de dos medias poblacionales, etc.).

Después de leer los capítulos 14 y 15 estarás mucho mejor preparado para cuando tengas que decidir a qué grupo debes orientar las acciones de marketing de un producto, qué marca de neumáticos dura más, qué eficacia tiene un determinado plan de adelgazamiento, y otras cuestiones más importantes como, por ejemplo, qué procedimiento quirúrgico deberías elegir.

Correlación, regresión y tablas de contingencia

Uno de los objetivos más comunes de las investigaciones consiste en encontrar vínculos entre variables. Por ejemplo,

- ✓ ¿Qué hábitos de vida aumentan o disminuyen el riesgo de padecer cáncer?
- ✓ ¿Qué efectos secundarios se asocian a este nuevo fármaco?
- ✓ ¿Puedo reducir el colesterol tomando este nuevo suplemento de hierbas?
- ✓ ¿Pasar mucho tiempo navegando por Internet te hace ganar peso?

Encontrar relaciones entre variables es lo que ayuda a la comunidad médica a diseñar mejores fármacos y tratamientos, proporciona a los expertos en marketing información sobre quién es más probable que compre sus productos, y permite a los políticos saber a quién deben dirigir sus argumentos a favor y en contra de determinadas políticas.



La búsqueda de relaciones entre variables genera muchos resultados estadísticos, pero ¿sabrías discriminar entre los que son correctos y los que no? Muchas decisiones importantes se basan en estos estudios, y por eso es importante saber qué normas deben cumplirse para que los resultados puedan considerarse fiables, sobre todo cuando se concluye la existencia de una relación causa-efecto.

En el capítulo 18 comento todos los pormenores y dificultades de crear representaciones gráficas a partir de dos variables numéricas (por ejemplo la dosis y la tensión arterial), encontrar e interpretar una *correlación* (la fuerza y la dirección de la relación lineal entre x e y), hallar la ecuación de una línea recta que se corresponda lo mejor posible con los datos

(y cuándo es apropiado hacer tal cosa), y cómo utilizar esos resultados para realizar predicciones sobre una variable a partir de otra variable (lo que se llama *regresión*). También te ofrezco herramientas para averiguar si una recta se corresponde o no con los resultados, y qué conclusiones puedes extraer (y las que no deberías extraer) en las situaciones en que sí se corresponde.

En el capítulo 19 comento con detalle varios métodos para encontrar y describir relaciones entre dos variables categóricas (por ejemplo, el número de dosis diarias y la presencia o ausencia de náuseas). También explico la manera de recopilar y organizar datos en *tablas de contingencia* (donde los valores posibles de una variable forman las filas y los valores posibles de la otra variable forman las columnas), interpretar los resultados, analizar los datos de dichas tablas para encontrar relaciones, y verificar la independencia. Asimismo, igual que hago en todo el libro, te doy estrategias para valorar de forma crítica si los resultados de estos tipos de análisis son fiables.

Extraer conclusiones fiables



Para realizar análisis estadísticos, los investigadores utilizan aplicaciones informáticas que dependen de fórmulas. Sin embargo, las fórmulas no saben si están siendo utilizadas en la forma debida, y no te avisan si los resultados son incorrectos. Por muchos cálculos que hagan, los ordenadores no te dirán nunca qué significan los resultados. Eres tú quien debe averiguarlo. A lo largo de este libro verás qué tipos de conclusiones puedes y no puedes extraer después de realizar el análisis. Los siguientes apartados son una introducción al arte de extraer conclusiones correctas.

Detectar resultados exagerados

Uno de los errores más comunes a la hora de extraer conclusiones consiste en exagerar los resultados o extrapolarlos a un grupo más grande que el representado por el estudio. Por ejemplo, un profesor quiere saber qué anuncios gustaron más a los espectadores de la final de la Copa del Mundo de fútbol, así que reúne a 100 alumnos suyos y les pide que puntúen los anuncios a medida que se los muestra. De esta forma crea un ranking y concluye que los cinco anuncios que ocupan los primeros puestos fueron los preferidos de todas las personas que vieron la final por televisión. Pero la verdad es que únicamente sabe qué anuncios gustaron más a sus alumnos. Como no estudió a ningún otro grupo, no puede extraer conclusiones sobre todos los televidentes.

Cuestionar afirmaciones de causa-efecto

Un ejemplo de razonamiento incorrecto es cuando los investigadores observan que dos variables están relacionadas (por medio de un análisis como, por ejemplo, la regresión; más información en el apartado anterior “Correlación, regresión y tablas de contingencia”) y automáticamente llegan a la conclusión de que esas dos variables presentan una relación de causa-efecto.

Pongamos que un investigador realiza una encuesta sobre salud y descubre que las personas que toman vitamina C todos los días se resfrían menos que quienes no toman vitamina C a diario. En vista de los resultados, y utilizando esos datos como prueba, escribe un artículo y lanza un comunicado de prensa diciendo que la vitamina C previene los resfriados.

Bueno, aunque quizá sea cierto que la vitamina C previene los resfriados, este estudio no tiene base suficiente como para afirmar tal cosa. Se trata de un estudio de observación, lo que significa que el investigador no controlaba otros factores que podrían estar relacionados con la vitamina C y con los resfriados. Por ejemplo, puede que las personas que toman vitamina C todos los días se preocupen más por su salud, se laven las manos más a menudo, hagan más ejercicio físico y coman alimentos más sanos, todo lo cual podría contribuir a resfriarse menos veces.



A menos que hagas un experimento controlado, no puedes extraer una conclusión causa-efecto a partir de las relaciones que observes (anteriormente en este mismo capítulo he hablado de los experimentos con más detalle).

Conviértete en un detective, no en un escéptico

La estadística es mucho más que jugar con los números. Para comprender de verdad la estadística tienes que aprender a extraer conclusiones correctas a partir de los datos estudiados y ser suficientemente listo como para no creerte todo lo que oigas o leas hasta que descubras cómo se obtuvo la información, qué se hizo con ella y cómo se llegó a esas conclusiones. A lo largo del libro me refiero varias veces a esta cuestión, pero donde más a fondo la comento es en el capítulo 20, que contiene diez formas de convertirse en un sabueso de las estadísticas y ser capaz de detectar los errores comunes que cometen los investigadores y los medios de comunicación.

26 Parte I: Estadísticas básicas sobre la estadística



Si eres estudiante, en el capítulo 21 encontrarás varios consejos que te ayudarán a preparar bien el examen y a sacar mejor nota. La mayoría de ellos se basan en adquirir una perspectiva más amplia y comprender la manera correcta de abordar los problemas estadísticos.



Resulta muy fácil adoptar una actitud escéptica o cínica frente a la estadística, sobre todo cuando sabes lo que ocurre entre bastidores. No dejes que eso te ocurra a ti. Por ahí fuera circula mucha información correcta que puede repercutir en tu vida de forma muy positiva. Canaliza tu escepticismo marcándote dos metas personales:

- ✓ Convertirte en un consumidor bien informado de las estadísticas que ves a diario.
- ✓ Aumentar la seguridad de tu puesto de trabajo siendo la persona a quien recurrir para temas relacionados con la estadística, la que sabe cuándo y cómo ayudar a los demás y cuándo debe consultarse a un estadístico.

Leyendo y utilizando la información de este libro tendrás la certeza de tomar decisiones correctas sobre resultados estadísticos, realizarás tus propios estudios estadísticos de manera creíble y estarás preparado para llevar a cabo tu próximo proyecto de trabajo, valorar esos molestos anuncios de los políticos, ¡y sacar un 10 en tu próximo examen!