¡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

¡La fórmula del éxito!

Tomamos un tema de actualidad y de interés general, añadimos el nombre de un autor reconocido, montones de contenido útil y un formato fácil para el lector y a la vez divertido, y ahí tenemos un libro clásico de la colección *Para Dummies*.

Millones de lectores satisfechos en todo el mundo coinciden en afirmar que la colección Para Dummies ha revolucionado la forma de aproximarse al conocimiento mediante libros que ofrecen contenido serio y profundo con un toque de informalidad y en lenguaje sencillo.

Los libros de la colección Para Dummies están dirigidos a los lectores de todas las edades y niveles del conocimiento interesados en encontrar una manera profesional, directa y a la vez entretenida de aproximarse a la información que necesitan.

¡Entra a formar parte de la comunidad Dummies!

El sitio web de la colección *Para Dummies* está pensado para que tengas a mano toda la información que puedas necesitar sobre los libros publicados. Además, te permite conocer las últimas novedades antes de que se publiquen y acceder a muchos contenidos extra, por ejemplo, los audios de los libros de idiomas.

Desde nuestra página web, también puedes ponerte en contacto con nosotros para comentarnos todo lo que te apetezca, así como resolver tus dudas o consultas.

También puedes seguirnos en Facebook (www.facebook.com/paradummies), un espacio donde intercambiar impresiones con otros lectores de la colección, y en Twitter @ParaDummies, para conocer en todo momento las últimas noticias del mundo Para Dummies.

10 cosas divertidas que puedes hacer en www.paradummies.es, en nuestra página en Facebook y en Twitter @ParaDummies

1. Consultar la lista completa de libros *Para Dummies*.
2. Descubrir las novedades que vayan publicándose.
3. Ponerte en contacto con la editorial.
4. Suscribirte a la Newsletter de novedades editoriales.
5. Trabajar con los contenidos extra, como los audios de los libros de idiomas.
6. Ponerte en contacto con otros lectores para intercambiar opiniones.
7. Comprar otros libros de la colección.
8. Publicar tus propias fotos en la página de Facebook.
9. Conocer otros libros publicados por el Grupo Planeta.
10. Informarte sobre promociones, descuentos, presentaciones de libros, etc.

Descubre nuestros interesantes y divertidos vídeos
en nuestro canal de Youtube:
www.youtube.com/paradummies
¡Los libros Para Dummies también están disponibles

en e-book y en aplicación para iPad!

La autora

Deborah J. Rumsey es especialista en Didáctica de la Estadística, profesora auxiliar en el Departamento de Estadística de la Universidad de Ohio State y miembro de la Sociedad Estadounidense de Estadística. Ha ganado el Premio a la Enseñanza de la Universidad de Kansas State y su nombre figura en el muro de la inspiración del centro de enseñanza secundaria de Burlington, Wisconsin, donde estudió. También es autora de *Statistics II For Dummies*, *Statistics Workbook For Dummies, Probability For Dummies* y *Statistics Essentials For Dummies*. Ha publicado numerosos artículos y ha presentado muchas ponencias sobre didáctica de la estadística. Es una de los organizadores del Congreso Bienal Estadounidense sobre Enseñanza de la Estadística (USCOTS, por sus siglas en inglés). Le gusta estar con su familia, salir de acampada y observar aves, conducir su tractor Kubota y animar al equipo de fútbol americano de la Universidad de Ohio State para que gane el próximo campeonato nacional.

Dedicatoria

A mi esposo Eric: eres mi sol y mi luna. A mi hijo Clint: te quiero hasta el infinito y más allá.

Agradecimientos de la autora

Mi más sincero agradecimiento a Lindsay Lefevere y Kathy Cox por darme la oportunidad de escribir varios libros de la colección *Para Dummies* de Wiley; a mis directores editoriales, Georgette Beatty, Corbin Collins y Tere Drenth por su constante apoyo y orientación; a Marjorie Bond del Monmouth College por ser mi editora técnica (¡otra vez!); a Paul Stephenson, que también intervino en la edición técnica, y a Caitie Copple y a Janet Dunn por su magnífica labor de corrección.

También le doy las gracias a Elizabeth Stasny, Joan Garfield, Kythrie Silva, Kit Kilen, Peg Steigerwald, Mike O'Leary, Tony Barkauskas, Ken Berk y Jim Higgins por la inspiración y la ayuda que me han brindado, y a toda mi familia por su amor incondicional y sus palabras de ánimo.

Introducción

Todos los días estás expuesto a una cantidad increíble de información estadística. Ya sabes a qué me refiero: gráficos, tablas y titulares que hablan de los resultados del último sondeo, encuesta, experimento o estudio científico. Este libro pretende que desarrolles y mejores tu capacidad de comprender, analizar y evaluar toda esa información, y que te diviertas por el camino. Aprenderás a descifrar y tomar decisiones importantes sobre resultados estadísticos (por ejemplo los resultados de los últimos estudios médicos) y tomarás conciencia de las formas en que pueden engañarte con las estadísticas. Además, verás cuál es la forma correcta de diseñar un estudio, recopilar datos, hacer cálculos y extraer conclusiones cuando te toque hacerlo a ti.

Esta obra también está diseñada para ayudar a todos los que estéis realizando un curso de introducción a la estadística y necesitéis algo de refuerzo. Aprenderéis los conceptos básicos de la estadística y reuniréis un montón de herramientas y trucos del oficio que os ayudarán a superar el escollo de los exámenes.

Además, está repleto de ejemplos reales de fuentes reales que tienen que ver con la vida cotidiana, desde los avances médicos recientes hasta estudios de delincuencia, tendencias de población y los últimos informes del gobierno de Estados Unidos. ¡Incluso he incluido un informe sobre los peores coches del milenio! Cuando termines el libro sabrás recopilar, representar y analizar datos de manera correcta y eficaz, y estarás preparado para realizar una evaluación crítica y tomar decisiones informadas sobre los últimos sondeos, encuestas, experimentos e informes con los que te bombardean a diario. ¡Hasta aprenderás a medir la temperatura oyendo el canto de un grillo!

También te reirás con algunos chistes sobre estadísticos (que a veces se toman a sí mismos demasiado en serio). Después de todo, si se tienen los conocimientos adecuados, no hace falta ser estadístico para comprender los fundamentos de la estadística.

Acerca de este libro

Este libro se aparta de los textos, referencias, libros de consulta y guías de estudio tradicionales sobre estadística por los siguientes motivos:

- ✔ Presenta explicaciones prácticas e intuitivas de los conceptos, ideas, técnicas, fórmulas y cálculos estadísticos que suelen aparecer en un curso de introducción a la estadística.
- ✔ Ofrece procedimientos paso a paso claros y concisos que te enseñan a resolver problemas estadísticos de forma intuitiva.
- ✔ Incluye interesantes ejemplos del mundo real que guardan relación con la vida cotidiana y el lugar de trabajo.
- ✔ Proporciona respuestas claras y sinceras a preguntas como "¿Qué significa esto en realidad?" y "¿Cuándo y cómo voy a utilizar esto otro?".

Convenciones utilizadas en este libro

Para avanzar sin tropiezos en la lectura debes conocer las siguientes convenciones:

- ✔ **Definición de tamaño muestral *(n)*:** Cuando hablo del tamaño muestral, me refiero al número final de personas que participaron y proporcionaron información para el estudio. Dicho de otro modo, *n* es el tamaño del conjunto de datos final.
- ✔ **Uso ambivalente de la palabra *estadístico*:** En algunas situaciones esta palabra se refiere a la persona experta en el tema de estudio o campo de investigación que nos ocupa, como cuando dices: "Los estadísticos son unos cabezas cuadradas". Otras veces utilizaré la palabra para referirme a una variable aleatoria o función de variables aleatorias, e incluso a su resultado concreto. Por ejemplo: "Los estadísticos básicos son la media y la desviación estándar".
- ✔ **Uso ambivalente de la palabra *estadística*:** En ocasiones la palabra nombra a la ciencia como tal, y otras veces, generalmente en plural, se refiere de manera genérica a los resultados finales del análisis estadístico tal y como se comunican al público.
- ✔ **Uso del término *desviación estándar*:** Cuando utilice el término *desviación estándar* me referiré a *s*, la desviación estándar de la muestra (si me refiero a la desviación estándar de la población, te avisaré).

Aquí tienes otras convenciones básicas que te facilitarán la lectura del libro:

- ✔ Utilizo la *cursiva* para introducir un nuevo término estadístico P y por las reglas tipográficas, también van en cursiva las variables (*x*, *y*, *n*, etc.).
- ✔ Si ves un término o una frase en **negrita** dentro de una lista, se trata de una palabra o frase clave.
- ✔ Las direcciones de Internet se indican en monofont.

Lo que no necesitas leer

Me gustaría pensar que no vas a saltarte ninguna parte del libro, pero también sé que eres una persona ocupada. Si quieres ahorrar tiempo, puedes saltarte las partes marcadas con el icono "Cuestiones técnicas" y los textos complementarios (los recuadros sombreados en gris que encontrarás a lo largo del libro). Toda esa información es interesante pero no imprescindible para tener un conocimiento básico de la estadística.

Algunas presuposiciones

No doy por sentado que tengas ningún tipo de experiencia anterior con la estadística, salvo por el hecho de que, como todo el mundo, estás expuesto a diario a un bombardeo continuo de estadísticas en forma de números, porcentajes, gráficos, resultados "estadísticamente significativos", estudios "científicos", sondeos, encuestas, experimentos, etc.

Lo que sí supongo es que sabes realizar algunas operaciones matemáticas elementales y entiendes la notación básica que se utiliza en álgebra, por ejemplo las variables *x* e *y*, el signo sumatorio, calcular raíces cuadradas y potencias, etc. Si necesitas refrescar tus conocimientos de álgebra, puedes leer *Matemáticas cotidianas para Dummies*, de Charles Seiter, adaptado para su publicación en España por Marc Meléndez.

Pero tampoco quiero engañarte: la estadística obliga a hacer algunos cálculos, y por eso este libro contiene fórmulas y ecuaciones. De todos modos, no hay de qué preocuparse. Te guiaré paso a paso por todos los cálculos que sean necesarios y te daré ejemplos para que practiques a medida que avanzas con el libro. Así te familiarizarás con las operaciones y no tendrás ningún problema.

Cómo está organizado el libro

Este libro se divide en cinco partes que tratan sobre las principales áreas de la estadística básica, más una parte final donde encontrarás unos decálogos la mar de útiles y divertidos. Cada parte, a su vez, consta de varios capítulos que desglosan esas áreas de la estadística en parcelas que puedan entenderse.

Parte I: Estadísticas básicas sobre la estadística

Esta parte te ayuda a tomar conciencia de la cantidad y la calidad de las estadísticas que encuentras en tu lugar de trabajo y en tu vida cotidiana. Descubrirás que buena parte de la información estadística es incorrecta, a veces por accidente y otras veces de manera intencionada. Como primer paso para ser un entendido en la materia, conocerás algunas de las herramientas del oficio, obtendrás una visión general de la estadística como proceso orientado a obtener e interpretar información, y aprenderás algunos términos especializados.

Parte II: Nociones sobre cálculo numérico

Esta parte te ayudará a familiarizarte con la elaboración, interpretación y evaluación de presentaciones de datos (también llamadas gráficos, diagramas, etc.). También aprenderás a resumir y examinar datos calculando y combinando varios estadísticos frecuentemente utilizados y otros que quizá no conozcas todavía.

Parte III: Distribuciones y el teorema del límite central

En esta parte conocerás a fondo las tres distribuciones estadísticas más frecuentes: la distribución binomial, la distribución normal (y la normal estándar, también llamada distribución *Z*), y la distribución *t*. Descubrirás las características de cada distribución y cómo encontrar e interpretar probabilidades, percentiles, medias y desviaciones estándares. También verás medidas de situación relativa (como los percentiles).

Por último, averiguarás cómo miden los estadísticos la variabilidad de las muestras y por qué es tan importante medir la precisión de los resultados de la muestra. Y también te explicaré con pelos y señales lo que algunos estadísticos llaman "la joya de la Corona de la estadística": el teorema del límite central (TLC). Yo no suelo utilizar este lenguaje tan florido para referirme al TLC, no vaya a ser que se me asusten los alumnos. En cuanto a los sentimientos que despierta en ellos el TLC, lo dejo a tu imaginación.

Parte IV: Estimaciones aproximadas y formulación de hipótesis fiables

Esta parte trata sobre los dos métodos que hay para generalizar los resultados de una muestra con el propósito de extraer conclusiones sobre una población entera (lo que los estadísticos llaman *inferencia estadística*). Estos dos métodos son los intervalos de confianza y el contraste de hipótesis.

En esta parte utilizaremos los intervalos de confianza para obtener estimaciones válidas para una o dos medias o proporciones poblacionales, o para la diferencia entre ellas (por ejemplo, el número medio de horas semanales que los adolescentes ven la televisión o el porcentaje de hombres y de mujeres estadounidenses que toman medicamentos para la artritis todos los días). Adquirirás los conocimientos básicos que te permitirán definir e interpretar intervalos de confianza y evaluar su corrección y credibilidad. Explorarás los factores que influyen en la amplitud de un intervalo de confianza (por ejemplo el tamaño de la muestra) y verás fórmulas, cálculos paso a paso y ejemplos para los intervalos de confianza más utilizados.

Los contrastes de hipótesis incluidos en esta parte enseñan la manera de utilizar los datos para verificar una aseveración sobre una o dos medias o proporciones poblacionales, o la diferencia entre ellas (por ejemplo, una empresa afirma que entrega los pedidos en el plazo de dos días como media: ¿es cierto?). Descubrirás qué procedimiento siguen (o deberían seguir) los investigadores para formular y verificar hipótesis y cómo puedes evaluar la precisión y credibilidad de sus resultados, y te daré instrucciones pormenorizadas y ejemplos con los que obtener e interpretar los resultados de los contrastes de hipótesis más utilizados.

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

Esta parte ofrece una visión general de las encuestas, los experimentos y los estudios de observación. Aprenderás para qué sirven esos estudios, cómo se llevan a cabo, cuáles son sus limitaciones y cómo deben evaluarse para determinar si los resultados son fiables.

También te explicaré con todo detalle la forma de analizar pares de variables numéricas y variables categóricas para encontrar relaciones, lo cual constituye el objeto de un gran número de estudios. Para los pares de variables categóricas hay que crear tablas de contingencia y encontrar probabilidades y distribuciones conjuntas, condicionales y marginales. Si se encuentra una relación de dependencia, se describe la naturaleza de dicha relación utilizando probabilidades. Para las variables numéricas se crean diagramas de dispersión, se encuentra y se interpreta la correlación existente, se realizan análisis de regresión, se estudia el ajuste de la recta de regresión y el impacto de los valores atípicos, se describe la relación por medio de la pendiente y se utiliza la recta para hacer predicciones. ¡Y todo eso en un día!

Parte VI: Los decálogos

En esta parte breve y sencilla encontrarás diez consejos para valorar datos estadísticos y para detectar estudios y resultados sospechosos, así como diez maneras seguras de sacar mejor nota en un examen de estadística.

Algunos cálculos estadísticos requieren el uso de tablas estadísticas. En el apéndice he incluido todas las que necesitarás para este libro. Estas tablas son la tabla *Z* (para la distribución normal estándar, también llamada distribución *Z*), la tabla *t* (para la distribución *t*) y la tabla binomial (para, lo has adivinado, la distribución binomial). Encontrarás instrucciones y ejemplos de uso de estas tres tablas en los correspondientes apartados de esta obra.

Iconos utilizados en este libro

Los iconos llaman la atención sobre determinadas cuestiones que aparecen de forma recurrente a lo largo del libro. Éstos son los iconos usados y su significado:

Este icono hace referencia a consejos, ideas o atajos que puedes utilizar para ahorrar tiempo. También señala otras maneras de enfocar un concepto en particular.

Este icono está reservado para algunas ideas concretas que espero que recuerdes mucho tiempo después de haber terminado el libro.

Este icono se refiere a maneras concretas en que los investigadores o los medios de comunicación pueden engañarte con las estadísticas, y te explica qué puedes hacer al respecto; además, señala posibles problemas y dificultades que debes tener en cuenta en los exámenes.

Busca este icono si tienes un interés especial en comprender los aspectos más técnicos de determinadas cuestiones estadísticas. Puedes saltártelo si no quieres entrar en tanto detalle.

Lo que tienes por delante

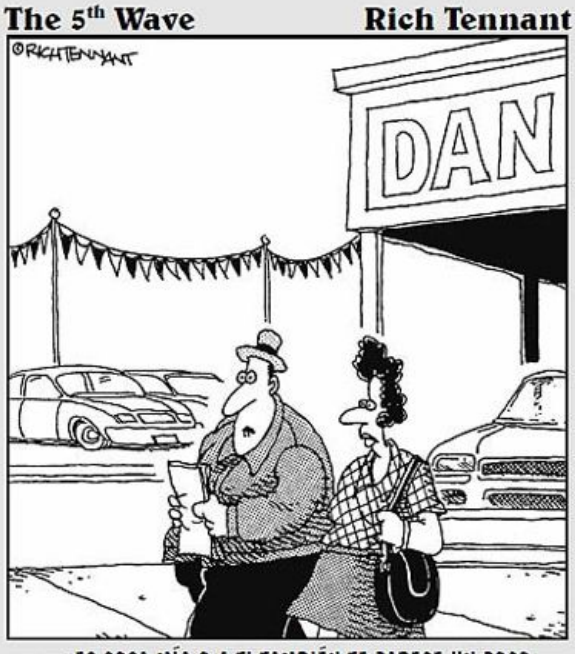
Este libro está escrito de forma que puedes comenzar por cualquier sitio e igualmente entender lo que pone. Por lo tanto, puedes echar un vistazo al sumario o al índice, buscar la información que te interese e ir directamente a la página en cuestión. De todos modos, si buscas información sobre un tema en particular y estás deseando sumergirte en él, aquí tienes algunas indicaciones:

- ✔ Para aprender a realizar e interpretar gráficos, diagramas, medias, medianas y similares, ve a la parte II.
- ✔ Si buscas información sobre la distribución normal, *Z*, *t* o binomial o sobre el teorema del límite central, mira en la parte III.
- ✔ Si quieres un monográfico sobre intervalos de confianza y contrastes de hipótesis de todo tipo, pasa a la parte IV.
- ✔ Para saber más sobre experimentos, regresiones y tablas de contingencia, busca en la parte V.

Si no estás seguro de por dónde empezar, lee primero el capítulo 1 para obtener una visión general y continúa luego con el resto del libro. ¡Que te diviertas leyendo!

Parte I

Estadísticas básicas sobre la estadística



—¿ES COSA MÍA O A TI TAMBIÉN TE PARECE UN POCO ESCASA ESA ESTADÍSTICA DE "SATISFACCIÓN AL 50%"?

En esta parte…

Cada vez que enciendes el televisor o abres un periódico te expones a un bombardeo de cifras, diagramas, gráficos y resultados estadísticos. La encuesta del día, el último avance en medicina... sólo ves números y más números. Sin embargo, buena parte de la información estadística que te obligan a consumir es errónea, unas veces por accidente y otras de manera intencionada. ¿Cómo puedes saber qué datos son ciertos? Haciendo una intensa labor de investigación.

Esta parte ayuda a despertar al sabueso de las estadísticas que hay en ti, para lo cual se analiza el modo en que las estadísticas influyen en tu vida cotidiana y en tu trabajo, hasta qué punto es incorrecta gran parte de la información que te dan y qué puedes hacer tú al respecto. Además esta parte te ayuda a familiarizarte con algunos términos especializados que te resultarán útiles más adelante.

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 repuestos 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 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 responder, 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 experimento 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 hacerte 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 *categóricos* (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.

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!

Capítulo 2

La estadística de la vida cotidiana

En este capítulo

- Plantearse preguntas sobre las estadísticas que aparecen en la vida cotidiana
- Entender las estadísticas propias del lugar de trabajo

La sociedad actual está inundada de cifras. Allá donde mires te encuentras con números y más números, desde carteles con las estadísticas de puntualidad de una determinada línea aérea hasta programas de deportes donde se comentan la probabilidad de victoria de los distintos equipos en la siguiente jornada. Los programas informativos están repletos de noticias sobre la tasa de delincuencia, la esperanza de vida de los adictos a la comida basura y el índice de popularidad del presidente del gobierno. En un día normal puedes encontrarte con cinco, diez o hasta veinte datos estadísticos distintos (y muchos más si es día de elecciones). Basta con leer de cabo a rabo el periódico de un domingo para tropezar con cientos de estadísticas en reportajes, anuncios y artículos que hablan de todo tipo de cosas, desde la sopa (¿cuántos litros consume al año una persona normal?) hasta los frutos secos (se sabe que las almendras son buenas para la salud, pero ¿ocurre lo mismo con otros frutos secos?).

En este capítulo hablo de las estadísticas que sueles encontrar en tu vida cotidiana y te cuento cómo se presentan al público en general. Cuando hayas terminado de leer el capítulo serás consciente de cuántas veces los medios de comunicación te lanzan estadísticas a la cara y cómo es de importante ser capaz de desentrañar el significado de esas cifras. Te guste o no, las estadísticas son una parte importante de tu vida. De manera que, si no puedes vencerlas, únete a ellas. Y si no quieres unirte a ellas, al menos intenta entenderlas.

Estadísticas en los medios de comunicación: ¿más preguntas que respuestas?

Abre un periódico y busca ejemplos de artículos y noticias donde te presenten cifras de cualquier tipo. Al poco rato te saldrán los números por las orejas. A los lectores les llueven resultados de estudios, anuncios de grandes avances científicos, informes estadísticos, previsiones, proyecciones, diagramas, gráficos y resúmenes. Es increíble la cantidad de estadísticas que aparecen en los medios. Puede que no seas consciente siquiera de cuántas veces te tropiezas con cifras a lo largo del día.

En este apartado veremos, a modo de ejemplo, unas pocas noticias sacadas de un periódico que leí el domingo pasado. Cuando veas la cantidad de veces que las noticias hablan de estadísticas sin proporcionarte toda la información que necesitas, a lo mejor empiezas a ponerte nervioso y a preguntarte qué es cierto y qué no lo es. ¡Tranquilo! Para eso tienes este libro, para ayudarte a distinguir entre la información buena y la mala (los capítulos de la parte II son un magnífico punto de partida).

Problemas con las palomitas

El primero de los artículos que contenía cifras se titulaba “Una fábrica de palomitas de maíz se enfrenta a una investigación de sanidad”, con el subtítulo “Los trabajadores enfermos dicen que los aromatizantes utilizados les dañaron los pulmones”. El artículo contaba que los centros de epidemiología habían expresado su preocupación por una posible relación entre la exposición a las sustancias químicas presentes en los aromatizantes de las palomitas de microondas y algunos casos de bronquiolitis obliterante. Ocho trabajadores de una misma fábrica habían contraído esta enfermedad y cuatro de ellos estaban a la espera de un trasplante pulmonar.

Según el artículo, había constancia de casos similares en otras fábricas de palomitas. Bueno, a lo mejor te estás preguntando qué pasa con las personas que comen palomitas de microondas. Según el artículo, los centros de epidemiología “no encontraron motivos para pensar que las personas que comen palomitas de microondas tengan algo que temer” (permaneced atentos). El siguiente paso consiste en evaluar más a fondo a los trabajadores por medio de encuestas para determinar su estado de salud y posibles exposiciones a los mencionados productos químicos, pruebas de capacidad pulmonar y espirometrías. La pregunta es la siguiente: ¿cuántos casos de esta enfermedad pulmonar constituyen una pauta verdadera y no mera casualidad o una anomalía estadística? (Encontrarás más información al respecto en el capítulo 14.)

La amenaza vírica

El segundo artículo hablaba sobre un reciente ciberataque: un virus gusano se abrió paso por Internet y logró ralentizar la navegación y la entrega de correos electrónicos en todo el mundo. ¿Cuántos ordenadores se vieron afectados? Los expertos citados en el artículo decían que había 39.000 ordenadores infectados, y que éstos a su vez afectaron a otros cientos de miles de sistemas.

Preguntas: ¿cómo llegaron los expertos a esa cifra? ¿De verdad examinaron todos los ordenadores para ver si estaban afectados? El hecho de que el artículo se escribiera menos de veinticuatro horas después del ataque apunta a que esa cifra es una conjetura. Entonces, ¿por qué decir 39.000 y no 40.000 (para que se notara menos que era una suposición)? Para saber más sobre estimaciones aproximadas (y cómo evaluar las cifras de otra persona), consulta el capítulo 13.

Entender los accidentes

Más adelante en el periódico alertaban sobre el elevado número de motociclistas fallecidos en accidentes de tráfico. Los expertos aseguraban

que la *tasa de mortalidad* de los motociclistas estadounidenses (el número de motociclistas fallecidos en accidente de tráfico por cada 100.000 vehículos matriculados) lleva tiempo aumentando, según un informe de la Administración Nacional Estadounidense para la Seguridad del Tráfico en Autopistas (NHTSA). En el artículo se mencionaban muchas causas posibles de este incremento de la mortalidad entre motociclistas, entre ellas la edad, sexo, cilindrada del motor, consumo de alcohol, conducir sin permiso y no usar el casco reglamentario. El reportaje era muy exhaustivo e incluía varias tablas y gráficos con los siguientes títulos:

> ✔ Motociclistas fallecidos y heridos, y tasas de mortalidad y de lesión por año, por número de vehículos matriculados y por millones de kilómetros recorridos.
>
> ✔ Número de motociclistas fallecidos en cada estado, uso de casco y contenido de alcohol en sangre.
>
> ✔ Tasa de mortalidad de pasajeros por tipo de vehículo (motocicletas, vehículos de transporte de pasajeros, camionetas), por cada 10.000 vehículos matriculados y por cada 100 millones de kilómetros recorridos.
>
> ✔ Número de motociclistas fallecidos según grupos de edad.
>
> ✔ Número de motociclistas fallecidos según cilindrada del motor.
>
> ✔ Historial de sanciones previas de los conductores implicados en accidentes mortales según tipo de vehículo (se incluían sanciones por conducir bajo los efectos del alcohol o las drogas, sanciones por exceso de velocidad y suspensiones o revocaciones del permiso de conducción).
>
> ✔ Porcentaje de motociclistas ebrios fallecidos en accidente de tráfico según hora del día, para accidentes de un solo vehículo, accidentes de varios vehículos y total de accidentes.

Este artículo era muy informativo y proporcionaba montones de datos sobre los fallecimientos y lesiones resultantes de accidentes de tráfico en Estados Unidos. Sin embargo, semejante avalancha de tablas, gráficos, tasas, cifras y conclusiones puede llegar a ser abrumadora e impedir que el lector se haga una idea clara de la situación. Con un poco de práctica, y con la ayuda de la parte II, te será más fácil entender los gráficos, tablas y diagramas y todos los estadísticos que los acompañan. Por ejemplo, es fácil confundirse entre tasas y cifras absolutas (no es lo mismo la tasa de mortalidad que el número de fallecimientos). Como comento en el capítulo 3, las cifras absolutas pueden resultar equívocas si se utilizan cuando sería más apropiado utilizar tasas.

A vueltas con las negligencias médicas

Más adelante en el mismo periódico había un artículo donde se comentaba un estudio reciente sobre los seguros de responsabilidad por negligencia médica en Estados Unidos. Las negligencias médicas afectan a la gente porque encarecen los honorarios de los médicos y además impiden a los enfermos obtener la asistencia médica que necesitan. En el artículo se decía que uno de cada cinco médicos del estado de Georgia había dejado de realizar procedimientos arriesgados (por ejemplo, asistir en partos) debido al fuerte incremento de precio de los seguros de responsabilidad por negligencia médica. Esta circunstancia se calificaba de epidemia nacional y crisis del sistema sanitario para todo el país. El artículo incluía algunos detalles del estudio y aseguraba que, de los 2.200 médicos encuestados en Georgia, 2.800 (según ellos, el 18% aproximadamente) iban a dejar de realizar procedimientos de alto riesgo.

¡Un momento! Esto está mal. ¿Son 2.200 médicos, hay 2.800 que no realizan esos procedimientos y se supone que representan el 18%? ¡Eso es imposible! Si el numerador de una fracción es más alto que el denominador, el resultado no puede ser inferior al 100%, ¿no? Éste es uno de los muchos errores que cometen los medios de comunicación cuando mencionan estadísticas. Entonces, ¿cuál es el porcentaje real? Es imposible saberlo a partir de la información del artículo. En el capítulo 5 explico a fondo los pormenores del cálculo estadístico, para que sepas lo que debes buscar y si hay algún error.

El problema de perder terreno

En el mismo periódico había un artículo sobre el alcance del desarrollo urbano y la especulación en Estados Unidos. Desde luego, siempre es interesante saber si van a construir más viviendas en tu zona. El artículo contenía estadísticas sobre el número de hectáreas de terreno agrícola que se urbanizan cada año. Para ilustrar mejor cuánto terreno se está perdiendo, la extensión se expresaba también como campos de fútbol. En este ejemplo concreto, los expertos decían que el estado de Ohio está perdiendo 61.000 hectáreas de suelo agrícola cada año, que son 610 kilómetros cuadrados o 115.385 campos de fútbol. ¿De dónde salen estos números y hasta qué punto son precisos? ¿De verdad sirve de algo pensar en el número de campos de fútbol a los que equivale la pérdida de terreno? En el capítulo 16 me refiero con más detalle a la precisión de los datos recopilados.

Los colegios en el punto de mira

La siguiente noticia hablaba de las competencias de los escolares estadounidenses, concretamente si asistir a clases extra les ayuda a mejorar las notas. El artículo decía que el 81,3% de los alumnos de un distrito en particular que asistieron a clases extra habían aprobado el examen de competencia escrita, mientras que sólo el 71,7% de los que no hicieron esas clases consiguieron aprobar. De acuerdo, pero ¿esa diferencia justifica los 386.000 dólares que cuestan las clases? ¿Y qué se hace en esas clases para conseguir una mejora? ¿Y si los alumnos dedican todo ese tiempo simplemente a preparar los exámenes en lugar de aprender más cosas sobre la escritura en general? Y la gran pregunta: ¿los alumnos que asistieron a esas clases extra lo hicieron por voluntad propia y, por tanto, estaban más motivados que un alumno normal para intentar mejorar sus notas de examen? En el artículo no lo decían.

Encuestas a tutiplén

Las encuestas y los sondeos son dos de los mecanismos más visibles utilizados por los medios de comunicación para captar tu atención. Parece que todo el mundo quiere hacer una encuesta,incluidos los directores de marketing, las compañías de seguros, las cadenas de televisión y hasta los alumnos de instituto. Aquí tienes unos cuantos resultados de encuestas que aparecen en las noticias de hoy:

En vista del progresivo envejecimiento de la población activa estadounidense, las empresas buscan nuevos directivos (¿cómo saben que la población activa estadounidense está envejeciendo? Y en tal caso, ¿cuánto está envejeciendo?). Según revela una reciente encuesta, casi el 67% de los directores de recursos humanos preguntados manifestaron que desde hace cinco años conceden cada vez más importancia a planificar la sucesión en los puestos de responsabilidad. La encuesta dice también que el 88% de los 210 encuestados generalmente promocionan a candidatos internos para ocupar cargos directivos. Pero ¿cuántos directores de recursos humanos no respondieron? ¿Y 210 encuestados son realmente suficientes para fundamentar una noticia que se publicó en la portada de la sección de economía? Lo creas o no, si te pones a buscar en cualquier periódico encontrarás numerosos ejemplos de noticias basadas en mucho menos de 210 participantes (para ser justos, no obstante, 210 puede ser un número razonable de sujetos en algunas situaciones. En el capítulo 16 me refiero con todo detalle a las cuestiones de qué tamaño de muestra y qué porcentaje de encuestados pueden considerarse suficientes).

Algunas encuestas se basan en intereses y tendencias actuales. Por ejemplo, según una encuesta reciente de la empresa Harris-Interactive, casi la mitad (el 47%) de los adolescentes estadounidenses aseguraron que sus vidas tocarían a su fin o serían mucho peores si les quitaran sus teléfonos móviles, y el 57% llegaron incluso a decir que los móviles son imprescindibles para su vida social. El estudio halló también que el 42% de los adolescentes dijeron ser capaces de mandar un mensaje de texto con los ojos vendados (¿cómo se comprueba tal cosa?). De todos modos, ten presente que el estudio no mencionaba qué porcentaje de adolescentes tienen móvil ni qué características demográficas presentan esos adolescentes en comparación con los que no tienen móvil. Y recuerda que los datos recopilados sobre cuestiones de este tipo no siempre son exactos, ya que los encuestados tienden a dar respuestas sesgadas (¿quién no querría decir que es capaz de mandar un mensaje de texto con los ojos vendados?). En el capítulo 16 encontrarás más información sobre cómo interpretar y valorar los resultados de encuestas.

Estudios de este tipo proliferan como setas, y la única forma de averiguar cuáles son fiables pasa por aprender a formular las preguntas adecuadas y ser capaz de criticar la calidad del estudio. ¡Todo eso forma parte de la estadística! Las buenas noticias son que bastan unas pocas preguntas aclaratorias para poder criticar un estudio estadístico y sus resultados. En el capítulo 17 te explico cómo se hace.

Los números del deporte

Las páginas deportivas probablemente sean la sección del periódico donde aparecen más números. Desde los resultados de los partidos hasta los porcentajes de victorias y derrotas y la posición relativa de cada equipo, las estadísticas especializadas que se utilizan en el mundo del deporte son tan variadas que abruman. Por ejemplo, las estadísticas de la liga de baloncesto se desglosan por equipos, cuartos y jugadores. Acerca de cada jugador te dicen los minutos jugados, los lanzamientos a canasta, los tiros libres, las asistencias, las faltas personales, las pérdidas de posesión, los tapones, los robos de balón y el total de puntos anotados.

¿Quién necesita toda esa información, aparte de las madres de los jugadores? Pues, al parecer, muchísima gente. Las estadísticas son un divertimento constante para los seguidores, una pesadilla para los jugadores y el tema de acaloradas discusiones entre aficionados que se las dan de entendidos.

Los simuladores de propiedad, por su parte, también generan un enorme volumen de negocio. Se trata de juegos en que los participantes actúan como propietarios y crean sus propios equipos a partir de jugadores reales de una liga profesional. Estos propietarios virtuales compiten entre sí y los resultados se deciden en función del rendimiento estadístico de los jugadores y equipos, medido a su vez con unas reglas que determina un comisario y un sistema de puntos establecido. Según la Fantasy Sports Trade Association, en Estados Unidos hay más de 30 millones de personas de doce años o más que utilizan estos juegos, y el volumen de negocio generado alcanza los 3.000-4.000 millones de dólares al año (y aquí también habría que saber cómo se han calculado esas cifras; ya ves que nunca termina uno de hacerse preguntas).

Información económica

En la sección de economía y negocios del periódico se encuentran estadísticas sobre el mercado de valores. Si te dicen que en una semana la Bolsa cayó 455 puntos, ¿dirías que es una bajada fuerte o suave? Para saberlo con certeza hay que calcular un porcentaje.

En las páginas de economía del periódico que suelo leer encontré información sobre los rendimientos más elevados de toda clase de certificados de depósito en Estados Unidos (por cierto, ¿cómo saben que esos rendimientos son los más elevados?). También encontré datos sobre préstamos a treinta años con tipo de interés fijo, préstamos a quince años con tipo de interés fijo, préstamos a un año con tipo de interés variable, préstamos para adquirir coches nuevos, préstamos para adquirir coches usados, préstamos con aval personal y préstamos de tu abuela (bueno, a lo mejor estoy exagerando, pero si mi abuela leyera estas estadísticas, seguro que me subía el interés).

Por último, vi muchos anuncios de tarjetas de crédito donde aparecían los tipos de interés, las tasas anuales y la duración del período de facturación. ¿Cómo puede uno comparar toda la información sobre inversiones, préstamos y tarjetas de crédito para tomar una buena decisión? ¿Qué estadísticas son las más importantes? La verdadera pregunta es la siguiente: ¿las cifras que aparecen en el periódico dicen toda la verdad o hace falta investigar más? En los capítulos 16 y 17 te ayudo a entender mejor todas esas cifras y a tomar decisiones sobre ellas.

El viaje de las cifras

No hay manera de escapar del aluvión de cifras, ni siquiera leyendo la sección de viajes. Por ejemplo, en esas páginas vi que la pregunta más frecuente que se formula al centro de información de la Administración de Estados Unidos para la Seguridad en el Transporte (que recibe alrededor de 2.000 llamadas telefónicas, 2.500 correos electrónicos y 200 cartas semanales de promedio —seguro que no querrías ser tú quien los contara, ¿verdad?) es: "¿Puedo llevar esto en el avión?". "Esto" puede referirse a cualquier cosa, desde un animal hasta un vestido de novia o un cubo gigante de palomitas (yo desaconsejaría llevar un cubo gigante de palomitas. Tendrías que ponerlo en posición horizontal en el compartimiento portaequipajes y, como las cosas se mueven durante el vuelo, es probable que se abriera la tapa, y entonces cuando fueras a coger el cubo al final del vuelo te caería una lluvia de palomitas a ti y a tus compañeros de asiento. Sí, yo fui testigo de eso en una ocasión).

El número de respuestas dadas en este caso plantea una interesante pregunta: ¿cuántos operadores se necesitan a distintas horas del día para atender todas las llamadas, correos electrónicos y cartas que se reciben? El primer paso sería estimar el número de llamadas que se espera recibir, sabiendo que una estimación errónea puede costarte dinero (si acaban siendo menos llamadas de las que pensabas) o mucha publicidad negativa (si son más de lo que creías). En el capítulo 13 hablo sobre este tipo de desafíos estadísticos.

Las estadísticas del sexo

Con todos los medios que tenemos a nuestro alcance hoy en día, resulta muy sencillo enterarse de las noticias y rumores más recientes, incluido el último estudio sobre vida sexual. En un artículo se decía que las personas casadas tienen 6,9 encuentros sexuales más cada año que las personas que han estado siempre solteras. Es una buena noticia para mí, supongo, pero ¿cómo pudo llegar alguien a esa cifra? En el artículo no ponía nada al respecto (puede que algunas estadísticas deban permanecer ocultas).

Si alguien hiciera una encuesta telefónica y pidiera a la gente que dedicara unos cuantos minutos de su tiempo a hablar sobre su vida sexual, ¿quiénes serían los que probablemente estarían más dispuestos a hablar del tema? ¿Y qué dirían cuando les preguntaran cuántas veces por semana mantienen relaciones sexuales? ¿Serían totalmente sinceros, te mandarían a paseo o exagerarían un poco? Las encuestas sobre temas personales pueden ser una fuente de sesgo y conducir a estadísticas erróneas. Pero entonces, ¿qué recomendarías a la gente que hiciera para saber más sobre una cuestión tan íntima como es el sexo? A veces investigar es más difícil de lo que parece (en el capítulo 16 hablo de los sesgos que aparecen al recopilar datos para ciertos tipos de encuestas).

Los misterios del parte meteorológico

Los partes meteorológicos son otra fuente de datos estadísticos, con previsiones de la temperatura máxima y mínima para el día siguiente (¿cómo deciden si serán 16 y no 15 grados centígrados?) y datos sobre el factor ultravioleta, la concentración de polen, el índice estándar de contaminación atmosférica y la cantidad y calidad del agua (¿de dónde salen todas esas cifras? ¿Es que toman muestras para todo? En ese caso, ¿cuántas muestras toman y en qué lugares?). Ahora mismo puedes saber qué tiempo hace en cualquier parte del mundo. Puedes ver una previsión meteorológica para los próximos tres días, para una semana, un mes o incluso un año. Los meteorólogos recopilan y registran miles y miles de datos sobre el tiempo todos los días. Esas cifras no sólo te ayudan a decidir si sales de casa con o sin paraguas, sino que permiten a los investigadores realizar predicciones a más largo plazo e incluso prever cambios climáticos a escala global.

Incluso con toda la información y las tecnologías que tienen a su disposición los meteorólogos hoy en día, ¿hasta qué punto son acertadas sus previsiones? A juzgar por el número de veces que llueve cuando se supone que iba a hacer sol, parece que todavía tienen mucho que mejorar. En cualquier caso, lo que sí demuestra esa enorme cantidad de datos es que el número de variables que influyen en el tiempo es abrumador, y no sólo para ti, sino también para los meteorólogos.

Los modelos estadísticos informatizados son muy importantes para realizar predicciones sobre grandes fenómenos meteorológicos como huracanes, terremotos y erupciones volcánicas. Los científicos todavía tienen que afinar un poco más para ser capaces de predecir un tornado antes de que empiece a formarse y decirte exactamente dónde y cuándo un huracán va a tocar tierra, pero no te quepa duda de que ése es su objetivo, y cada vez lo hacen mejor. En el capítulo 18 encontrarás más información sobre creación de modelos.

Reflexiones sobre las películas

En las páginas de espectáculos encontré varios anuncios de las películas que estaban en cartelera. Cada anuncio incluía frases textuales de algunos críticos: "¡Sensacional!", "La mejor aventura de todos los tiempos", "Para morirse de risa" o "Una de las diez mejores películas del año". ¿Tú haces caso a los críticos? ¿Cómo eliges las películas que vas a ver al cine? Los expertos aseguran que, si bien durante los primeros días la popularidad de una película puede verse afectada por las reseñas (buenas o malas) de los críticos, el boca a boca es lo que determina el resultado de taquilla a largo plazo.

Los estudios revelan también que, cuanto más espectacular es una película, más palomitas se venden. Sí, la industria del entretenimiento analiza incluso lo que comes en el cine. ¿De dónde sacan toda esa información y cómo influye en el tipo de películas que se hacen? Esto también forma parte de la estadística: diseñar y realizar estudios que ayuden a identificar un público y descubrir lo que les gusta, y luego utilizar esa información para orientar la creación de un producto. Así pues, la próxima vez que alguien con una carpeta te pregunte si tienes un minuto para responder a unas preguntas, piensa que tus respuestas cuentan.

Los horóscopos, en el punto de mira

Posiblemente leas los horóscopos, pero ¿te crees lo que dicen? ¿Deberías creerlo? ¿Es posible que alguien prediga el futuro con más acierto que el que puede atribuirse a la simple casualidad? Los estadísticos tienen una manera de descubrirlo, utilizando algo que han dado en llamar *contraste de hipótesis* (ver el capítulo 14). ¡Por el momento no han encontrado a nadie capaz de leer la mente, pero la gente todavía no se da por vencida!

Estadísticas en el trabajo

Ahora deja el periódico a un lado y piensa en el ajetreo de un día de trabajo cualquiera. Si eres contable, no cabe duda de que los números forman parte de tu vida cotidiana. Pero ¿qué pasa con las enfermeras, fotógrafos de estudio, gerentes de tiendas, periodistas, oficinistas o albañiles? ¿Las cifras influyen de algún modo en sus trabajos? Puedes apostar a que sí. En este apartado te ofrezco unos pocos ejemplos de cómo las estadísticas se cuelan en todos los lugares de trabajo.

No hace falta ser muy perspicaz para ver que la estadística está muy presente en nuestra vida personal y profesional. La clave está en comprender el significado de todos esos datos, saber cuáles son fiables y tomar decisiones bien fundamentadas a partir de la realidad que se oculta tras los números, para así saber hacer frente y acostumbrarte a las estadísticas de la vida cotidiana.

Recién nacidos e información

Elena trabaja de enfermera en el turno de noche de la unidad de obstetricia y ginecología de un hospital universitario, y todas las noches se ocupa de varias mujeres que acuden a dar a luz. Su enfermera jefe le ha dicho que al comenzar cada turno debe presentarse a las pacientes que le toquen, escribir su nombre en la pizarrita que hay en cada habitación y preguntarles a todas si tienen alguna pregunta. ¿Por qué? Pues porque pocos días después de que la madre vuelva a casa con su hijo recién nacido, el hospital la llamará por teléfono para preguntarle sobre la calidad de la asistencia recibida, si ha echado algo en falta, cómo se podría mejorar el servicio y la calidad de la asistencia, y qué podría hacer el personal para que ese hospital sea mejor que otros de la misma ciudad. Por ejemplo, las encuestas revelan que los pacientes que conocen el nombre de sus enfermeras se sienten más cómodos, hacen más preguntas y su experiencia en el hospital es mejor. Que en el futuro le den a Elena un aumento de sueldo depende de su capacidad para satisfacer las necesidades de las parturientas. En cualquier caso, no cabe duda de que el hospital se habrá preocupado de determinar otros factores que también influyen en la calidad de la asistencia médica, aparte de la interacción con las enfermeras. (En el capítulo 17 encontrarás información exhaustiva sobre los estudios médicos.)

Posar o no posar

Carolina ha empezado a trabajar hace poco como fotógrafa para un estudio situado en unos grandes almacenes. Uno de sus puntos fuertes son los retratos de bebés. A partir de los datos de venta recopilados a lo largo de los años, los responsables del estudio han observado que la gente compra más fotos que son resultado de un posado que fotos espontáneas. En consecuencia, animan a sus fotógrafos a que los hagan.

Una mujer entra con su hijo recién nacido y formula una petición especial: "¿Podría evitar que parezca un posado? Me gustan las fotos naturales". Si Carolina le responde: "Lo lamento pero no puedo. Mi futuro en este estudio depende de mi capacidad para retratar buenos posados", puedes estar seguro de que la madre rellenará el cuestionario sobre calidad del servicio nada más terminar la sesión, y no sólo para conseguir dos dólares de descuento en su próxima visita (suponiendo que haya una próxima visita). En lugar de eso, Carolina debería mostrar a su jefe la información contenida en el capítulo 16, donde hablo de la recogida de datos sobre satisfacción de los clientes.

Datos sobre pizzas

Pedro es el encargado de un establecimiento donde venden pizza en porciones. Uno de sus cometidos consiste en determinar cuántos empleados debe haber trabajando en un momento dado, cuántas pizzas hay que preparar de forma anticipada para satisfacer la demanda prevista, y cuánto queso hay que pedir y rallar, todo ello con el mínimo desperdicio posible de sueldos e ingredientes. Es medianoche de un viernes y el local está vacío. A Pedro le quedan cinco empleados y tiene cinco pizzas grandes que podría meter en el horno para obtener 40 porciones de cada una. ¿Debería mandar a casa a dos de sus empleados? ¿Debería meter más pizza en el horno o esperar a ver qué pasa?

Hace varias semanas que el dueño del establecimiento está llevando un seguimiento de la demanda, y por eso Pedro sabe que todos los viernes entre las diez y las doce de la noche hay poca afluencia de gente, pero alrededor de la medianoche el local empieza a abarrotarse y se mantiene lleno hasta las dos y media, cuando cierran. Por tanto, Pedro mantiene a todos los empleados, manda que metan una pizza en el horno cada treinta minutos a partir de la medianoche y se ve recompensado con una buena caja, muchos clientes satisfechos y un jefe contento. En el capítulo 13 encontrarás más información sobre cómo hacer buenas estimaciones utilizando la estadística.

Estadísticas en la oficina

José es auxiliar administrativo en una empresa informática. ¿Crees que en su trabajo también se cuelan las estadísticas? La respuesta es sí. En todas las oficinas hay un montón de personas que quieren conocer las respuestas a diversas preguntas, y necesitan a alguien que haga las cuentas, que les explique el significado de una información concreta, que averigüe si existen datos fidedignos sobre algo, o que simplemente

confirme si una cifra tiene sentido. Hay muchas cosas que deben saberse, desde los datos sobre satisfacción de los clientes hasta los cambios de inventario ocurridos durante el año, pasando por el porcentaje de tiempo que pasan los empleados escribiendo y leyendo correos electrónicos y el coste de aprovisionamiento de los últimos tres años. En todos los lugares de trabajo se manejan estadísticas, y el valor de José como empleado subiría muchos puntos si fuera la persona a quien todos recurren cuando tienen dudas. Todas las oficinas necesitan un estadístico en nómina, ¿por qué no serlo tú?

Capítulo 3

Tomar el control: tantos números y tan poco tiempo

En este capítulo

- Ver el mal uso de las estadísticas
- Comprender las repercusiones de unas estadísticas incorrectas

Con todas las estadísticas que vemos a diario, es normal que a veces nos sintamos abrumados y confundidos. En este capítulo te ofrezco una herramienta para hacer frente a este problema: ¡el escepticismo! No un escepticismo radical del estilo de "Ya no me creo nada", sino un escepticismo sano, como cuando te preguntas de dónde han sacado un dato o quieres tener más información antes de aceptar unos resultados. Para desarrollar un escepticismo sano tienes que comprender el funcionamiento de la cadena de información estadística.

Las estadísticas que ves en la televisión y en el periódico son el resultado de un proceso. En primer lugar, los investigadores que estudian un tema generan resultados; este grupo se compone de encuestadores, médicos, investigadores de marketing, investigadores del gobierno y otros científicos. Se los considera las *fuentes originales* de la información estadística.

Una vez que han obtenido unos resultados, esos investigadores, como es natural, quieren enseñárselos a la gente, de manera que emiten un comunicado de prensa o publican un artículo en una revista. Ahí es donde entran en juego los periodistas, considerados las *fuentes periodísticas* de la información. Los periodistas buscan comunicados de prensa interesantes y leen revistas especializadas para encontrar su próximo titular. Cuando los periodistas terminan de escribir sus noticias y reportajes, las estadísticas se comunican de inmediato al público en general a través de todo tipo de medios. Llegados a ese punto, la información ya está lista para ser absorbida por el tercer grupo: los *consumidores* de la información (o sea, tú). Tú y otros consumidores de información os enfrentáis a la tarea de escuchar y leer dicha información, analizarla y decidir sobre ella.

En cualquier etapa de este proceso de investigación, comunicación de resultados y consumo de información pueden cometerse errores, ya sea sin querer o queriendo. Las herramientas y estrategias que encontrarás en este capítulo te ayudarán a ser un buen detective.

Detectar errores, exageraciones y mentiras

Las estadísticas pueden estar equivocadas por muchas razones. En primer lugar, puede que alguien cometa un error sin querer. Eso le puede pasar a cualquiera, ¿no? Otras veces, sin embargo, el error no es tan inocente. En caliente, porque alguien cree ciegamente en algo y porque las cifras no terminan de sustentar la opinión que el investigador quiere defender, las estadísticas se alteran o, lo que es más común, se exageran, ya sea en cuanto a sus valores o en la forma en que se representan y se comentan.

Otro tipo de error es el *error de omisión*, que es el que se comete cuando se calla información que supondría una gran diferencia a la hora de interpretar la verdad que se oculta tras las cifras. Esa omisión dificulta la tarea de juzgar la corrección de un estudio, ya que no dispones de la información necesaria.

Incluso pueden darse casos en que las cifras sean totalmente inventadas y nadie más pueda repetirlas porque nunca fueron reales. En este apartado encontrarás consejos para detectar errores, exageraciones y mentiras, así como varios ejemplos de cada uno de los tipos de error con los que tú, como consumidor de información, puedes tropezar.

Comprobar las operaciones

Lo primero que deberías hacer al conocer una estadística o con el resultado de un estudio estadístico es preguntarte si esa cifra es correcta. ¡No des por sentado que lo es! Te sorprendería saber cuántos errores aritméticos se cometen al recopilar, resumir, comunicar o interpretar estadísticas.

Para detectar errores aritméticos u omisiones en estadísticas:

- ✔ **Comprueba que los totales cuadran.** Dicho de otro modo, asegúrate de que los porcentajes del gráfico de sectores suman 100 (o casi 100, debido al redondeo), y que la suma de personas de las distintas categorías coincide con el número total de personas encuestadas.
- ✔ **Revisa incluso los cálculos más elementales.**
- ✔ **Busca siempre un total para poder ver los resultados con perspectiva.** No hagas caso de los resultados basados en muestras muy pequeñas.

✔ **Comprueba si las proyecciones son razonables.** Por ejemplo, si se afirma que cada minuto mueren tres personas debido a una determinada enfermedad, eso son 1,5 millones de fallecimientos al año. Según cuál sea la enfermedad, puede que esa cifra no sea razonable.

Descubrir estadísticas engañosas

El engaño estadístico más común, con diferencia, consiste en exagerar la verdad de manera sutil pero efectiva. Incluso si los números cuadran, las propias estadísticas pueden ser engañosas cuando exageran los hechos. Las estadísticas engañosas son más difíciles de detectar que los simples errores de cálculo, pero pueden tener una fuerte repercusión sobre la sociedad. Además, por desgracia, son muy frecuentes.

Analizar los debates estadísticos

Los datos sobre delincuencia son un buen ejemplo de cómo la estadística se utiliza para mostrar las dos caras de una realidad, aunque a la postre sólo una de ellas sea correcta. La delincuencia es un tema recurrente en los debates políticos. Uno de los candidatos (generalmente el gobernante) asegura que la delincuencia ha descendido durante su mandato, mientras que el opositor dice que ha aumentado (así tiene algo que reprochar al equipo de gobierno). ¿Cómo es posible que dos candidatos lleguen a conclusiones tan diferentes a partir del mismo conjunto de datos? Pues ocurre que, según la manera en que se mida la delincuencia, es factible obtener ambos resultados.

En la tabla 3-1 se registran la población de Estados Unidos entre 1998 y 2008, el número de delitos denunciados y la *tasa* de delincuencia (delitos cometidos por cada 100.000 personas), que se calculan dividiendo el número de delitos por el tamaño de la población y multiplicando por 100.000.

Tabla 3-1. Número de delitos, tamaño estimado de la población y tasa de delincuencia en Estados Unidos

Año	*N.º de delitos*	*Tamaño de la población*	*Tasa de delincuencia por 100.000 personas*
1998	12.475.634	270.296.000	4.615,5
1999	11.634.378	272.690.813	4.266,5
2000	11.608.072	281.421.906	4.124,8
2001	11.876.669	285.317.559	4.162,6
2002	11.878.954	287.973.924	4.125,0
2003	11.826.538	290.690.788	4.068,4
2004	11.679.474	293.656.842	3.977,3
2005	11.565.499	296.507.061	3.900,6
2006	11.401.511	299.398.484	3.808,1
2007	11.251.828	301.621.157	3.730,5
2008	11.149.927	304.059.784	3.667,0

Fuente: Encuesta nacional sobre víctimas de la delincuencia en Estados Unidos.

Ahora compara el número de delitos y las tasas de delincuencia para 2001 y 2002. En la columna 2 puedes ver que en 2002 se cometieron 285 delitos más que en 2001 (11.878.954-11.876.669). Esto representa un incremento del 0,019% (se calcula dividiendo la diferencia, 2.285, por el número de delitos cometidos en 2001, 11.876.669). Observa que entre 2001 y 2002 también aumentó el tamaño de la población (columna 3), concretamente en 2.656.365 personas (287.973.924-285.317.559), que es el 0,931% (dividiendo esta diferencia por el tamaño de la población en 2001). Sin embargo, en la columna 4 vemos que la *tasa* de delincuencia disminuyó, pasando de 4.162,6 (por cada 100.000 personas) en 2001 a 4.125 (por cada 100.000 personas) en 2002. ¿Cómo es posible que disminuyera la tasa de delincuencia? Aunque el número de delitos y el número de personas aumentaron, el número de delitos lo hizo en menor medida que el tamaño de la población (el 0,019% frente al 0,931%).

Entonces, ¿cómo habría que comunicar esta tendencia? ¿La delincuencia aumentó o disminuyó de 2001 a 2002? Si tenemos en cuenta la tasa de delincuencia (que es una medida más exacta), podemos concluir que la delincuencia disminuyó durante ese año. Pero entonces es cuando entra en escena el opositor, quien, para demostrar que el gobernante no cumplió con su cometido, se verá tentado a contemplar el número de delitos y afirmar que la delincuencia aumentó, creando así una controversia artificial y confusión (por no hablar de escepticismo) entre los votantes. Para que luego digan que las campañas electorales no son divertidas.

A fin de establecer unas condiciones objetivas cuando se quiere medir la frecuencia con que ocurre un determinado suceso, todos los números deben convertirse a porcentajes dividiendo por el total para obtener lo que los estadísticos llaman una *tasa*. Las tasas generalmente son preferibles a las cifras absolutas porque permiten realizar comparaciones ecuánimes cuando los totales son diferentes.

Estadísticas sobre tornados

¿En qué estado de Estados Unidos se forman más tornados? Depende de cómo lo enfoques. Si te limitas a contar el número de tornados ocurridos en un determinado año (el dato que suelen utilizar los medios de comunicación), el primero de la lista es el estado de Texas. Sin embargo, la cosa no es tan sencilla. Texas es el segundo estado más grande de Estados Unidos (después de Alaska) y además se encuentra en una zona del país llamada "Corredor de los tornados". O sea, que le llegan muchos tornados y encima ocupa una extensión enorme donde esos tornados pueden tocar tierra y moverse. Una comparación más ecuánime, y la forma en que lo enfocan los meteorólogos, consiste en considerar el número de tornados por cada 10.000 millas cuadradas (que son 25.900 kilómetros cuadrados). Si utilizamos este estadístico (según cuál sea la fuente), Florida pasa a ocupar el primer puesto, seguido de Oklahoma, Indiana, Iowa, Kansas, Delaware, Luisiana, Misisippi y Nebraska, y luego Texas en décima posición (estoy seguro de que no les importa ocupar un puesto bajo en este ranking; otra cosa muy distinta es la clasificación de sus equipos en la liga universitaria de fútbol americano).

Otras medidas estadísticas de tornados que se dan a conocer en Estados Unidos incluyen el estado con el porcentaje más alto de tornados devastadores (Tennessee), y la distancia total recorrida por tornados por cada 10.000 millas cuadradas (Misisippi). Observa que estas estadísticas se indican como *tasa* (tanto por unidad).

Antes de creerte unas estadísticas sobre "el XXX más elevado" o "el XXX más bajo", comprueba si han utilizado un método ecuánime para medir la variable y si hay otras estadísticas que también deberían tenerse en cuenta para saber toda la verdad, y asegúrate de que las unidades son adecuadas para establecer comparaciones equitativas.

Ojo con la escala

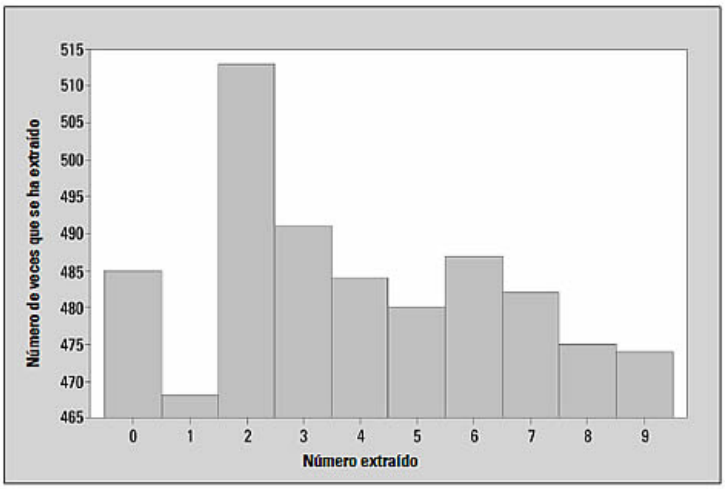
Los diagramas y los gráficos son útiles para comunicar los datos de forma rápida y clara. Por desgracia, muchas veces los diagramas y gráficos que acompañan a las estadísticas no se elaboran con la debida corrección o ecuanimidad. Uno de los elementos más importantes que debes tener en cuenta es la manera de escalar el diagrama o gráfico. La *escala* de un gráfico es la cantidad utilizada para representar cada una de las marcas en el eje del gráfico. Las marcas pueden indicar unidades, decenas, veintenas, centenas, etc. La escala influye mucho en el aspecto del gráfico o diagrama.

Por ejemplo, el organismo de loterías y apuestas del estado de Kansas habitualmente muestra los resultados recientes del sorteo Pick 3. Una de las estadísticas mostradas es el número de veces que cada número (del 0 al 9) sale escogido entre los tres números ganadores. La tabla 3-2 muestra el número de veces y el porcentaje de veces que se extrajo cada número en un total de 1.613 sorteos Pick 3 (se extrajeron 4.839 números individuales). Igual que antes, según cómo se contemplen los resultados podemos hacer que las estadísticas cuenten una historia muy diferente.

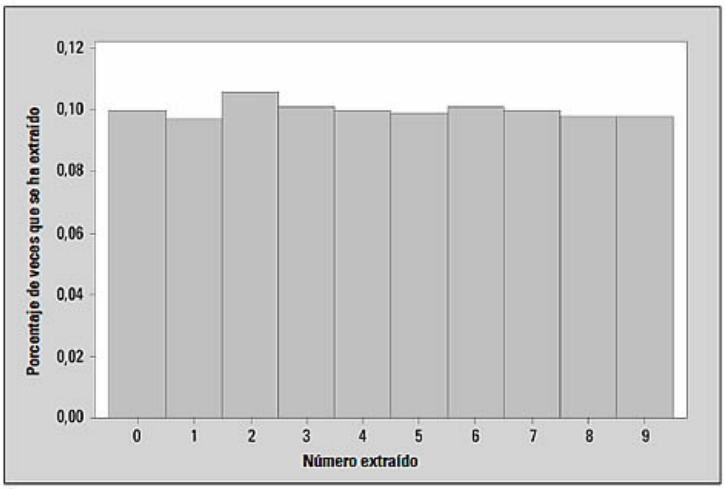
Tabla 3-2. Números extraídos en el sorteo Pick 3

Número extraído	*N.° de extracciones en 4.839 veces*	*Porcentaje de extracciones (n.° de extracciones / 4.839)*
0	485	10 %
1	468	9,7 %
2	513	10,6 %
3	491	10,1 %
4	484	10 %
5	480	9,9 %
6	487	10,1 %
7	482	10 %
8	475	9,8 %
9	474	9,8 %

En la figura 3-1a puedes ver la manera en que los organismos de loterías y apuestas suelen representar gráficamente resultados como los que aparecen en la tabla 3-2. Viendo el gráfico da la sensación de que el número 1 se extrae mucho menos a menudo (sólo 468 veces) que el número 2 (513 veces). La diferencia de altura entre estas dos barras parece muy grande, con lo que se exagera la diferencia en el número de veces que se han extraído estos dos números. Sin embargo, para ponerlo en perspectiva, la diferencia real es 513–468=45 de un total de 4.839 números extraídos. En porcentajes, la diferencia entre el número de veces que se extrae el número 1 y el número 2 es 45/4.839=0,009, es decir, solamente nueve décimas partes del 1%.



a



b

Figura 3-1: Gráficos de barras que muestran a) el número de veces que se ha extraído cada número, y b) el porcentaje de veces que se ha extraído cada número

¿A qué se debe que este gráfico exagere las diferencias? Existen dos razones. En primer lugar, el eje vertical, que representa el número de veces (o frecuencia) que se extrae cada número, aumenta de cinco en cinco. Para un total de 4.839 números extraídos, una diferencia de cinco parece significativa. Estirar la escala para que las diferencias parezcan mayores de lo que son en realidad es un truco utilizado a menudo para exagerar resultados. En segundo lugar, el gráfico empieza a contar en 465, no en 0. Sólo se muestra la parte superior de cada barra, lo que también contribuye a exagerar los resultados. En comparación, la figura 3-1b representa el porcentaje de veces que se ha extraído cada número. Normalmente la forma de un gráfico no cambiaría al pasar de cifras absolutas a porcentajes; sin embargo, este gráfico utiliza una escala más realista que la de la figura 3-1a (con incrementos del 2%) y además comienza en el 0, y ambos factores combinados hacen que las diferencias se vean tal y como son en realidad (muy pequeñas). Aburrido, ¿verdad?

Supongo que los de las loterías pensaron lo mismo. De hecho, a lo mejor utilizan la figura 3-1a en lugar de la figura 3-1b porque quieren que creas que los números se rigen por algún tipo de "magia" que nadie puede controlar. Al fin y al cabo, con eso se ganan la vida.

Tener en cuenta la escala de un diagrama o un gráfico puede ayudarte mucho a ver los resultados con la perspectiva adecuada. Estirar la escala o comenzar el eje *y* en el número más alto posible hace que las diferencias parezcan mayores, mientras que encoger la escala o comenzar el eje *y* en un valor mucho más bajo hace que las diferencias parezcan menores de lo que son en realidad.

Atención a las fuentes

Cuando examines los resultados de un estudio, comprueba de dónde procede la información. Los mejores resultados suelen publicarse en revistas acreditadas y bien conocidas por los expertos en ese campo. En el ámbito de la medicina, por ejemplo, el *Journal of the American Medical Association* (JAMA), el *New England Journal of Medicine*, *The Lancet*, y el *British Medical Journal* son algunas de las revistas especializadas que los investigadores utilizan para publicar resultados y enterarse de nuevos avances.

Ten siempre presente cuál es la fuente y quién ha financiado la investigación. Muchas empresas financian investigaciones y las utilizan para anunciar sus productos. Aunque esta práctica no es necesariamente censurable, en algunos casos puede existir un conflicto de intereses por parte de los investigadores que conduzca a resultados sesgados. Si los resultados son muy importantes para ti, entérate de si se ha realizado más de un estudio y, en tal caso, pide que te dejen revisarlos todos, no sólo aquellos cuyos resultados se hayan publicado en revistas o hayan aparecido en anuncios.

Fíjate en el tamaño de la muestra

El tamaño de la muestra no lo es todo, pero sí tiene mucho peso en las encuestas y estudios. Si el estudio se diseña y se lleva a cabo correctamente y los participantes se seleccionan de forma aleatoria (es decir, sin sesgo; en el capítulo 16 hablo con detalle sobre las muestras aleatorias), el tamaño de la muestra es un factor importante para determinar la precisión y fiabilidad de los resultados. (Para más información sobre el diseño y la realización de estudios, lee los capítulos 16 y 17.)

Muchas encuestas cuentan con un gran número de participantes, pero no siempre ocurre así con otros tipos de investigación, por ejemplo los experimentos controlados. Debido a la fuerte inversión de tiempo y dinero que requieren determinados tipos de investigación, algunos estudios se basan en un reducido número de participantes o productos. Los investigadores deben encontrar un equilibrio adecuado a la hora de determinar el tamaño de la muestra.

Los resultados menos fiables son los basados en *anécdotas*, es decir, historias que se refieren a un incidente aislado con la intención de influir en las corrientes de opinión. ¿Alguna vez le has dicho a alguien que no compre un producto porque a ti no te ha dado buen resultado? Recuerda que una anécdota es en realidad una muestra no aleatoria con un tamaño igual a uno.

Reflexiones sobre causa y efecto

Los titulares a menudo simplifican o deforman la información "real", sobre todo cuando las noticias se refieren a estadísticas y a los estudios que generaron esas estadísticas.

Un estudio realizado hace unos años evaluaba las filmaciones de 1.265 visitas de 59 médicos de familia y 6 cirujanos en los estados de Colorado y Oregón. El estudio halló que los médicos que no habían sido demandados por negligencia dedicaban una media de dieciocho minutos a cada paciente, mientras que los médicos que sí habían sido demandados invertían dieciséis minutos por visita. El estudio apareció en los medios con el titular siguiente: "Tratar bien a los enfermos evita demandas por negligencia médica". Lo que el estudio parecía indicar es que, si eres médico y pueden demandarte, basta con que dediques más tiempo a tus pacientes para estar fuera de peligro (¿por qué el periodista consideró que tratar bien a los pacientes es lo mismo que dedicarles más tiempo?).

En cualquier caso, ¿de verdad deberíamos creer que a un médico que ha sido demandado le basta con dedicar un par de minutos más a cada paciente para evitar nuevas demandas en el futuro? Quizá lo que el médico haga durante esos dos minutos cuente mucho más que el tiempo total dedicado a cada paciente. En el capítulo 18 puedes leer sobre los problemas de las relaciones causa-efecto entre variables.

Encontrar lo que querías encontrar

Quizá te hayas preguntado alguna vez cómo es posible que dos políticos se refieran a una misma cuestión y extraigan conclusiones opuestas, ambas basadas en "encuestas científicas". Incluso pequeñas diferencias en una encuesta pueden dar lugar a grandes diferencias en los resultados. (El tema de las encuestas se trata a fondo en el capítulo 16.)

Una causa frecuente de sesgo en los resultados de una encuesta es la manera de formular las preguntas. Aquí tienes tres preguntas diferentes que pretenden dilucidar una misma cosa: cuál es la opinión de la gente sobre la opción de veto de partidas específicas por parte del presidente de Estados Unidos:

- ✔ ¿Debería el presidente disponer de la opción de veto para eliminar las partidas que no sirvan de nada (sí, no, no sabe/no contesta)?
- ✔ ¿La opción de veto de partidas específicas otorga demasiado poder individual al presidente (sí, no, no sabe/no contesta)?
- ✔ ¿Qué opina usted de la opción de veto de partidas específicas por parte del presidente? Elija una respuesta entre el 1 y el 5, donde el 1 significa "totalmente en contra" y el 5 significa "totalmente a favor".

Las primeras dos preguntas son engañosas y conducen a resultados sesgados en sentidos opuestos. La tercera versión permitirá obtener resultados más exactos y, por tanto, averiguar qué piensa la gente en realidad. No obstante, no todas las encuestas se redactan con el propósito de encontrar la verdad; muchas se elaboran para apoyar un determinado punto de vista.

Los estudios demuestran que incluso pequeños cambios en la redacción de las preguntas influyen mucho en las respuestas de la gente, con lo que se obtienen resultados contradictorios al comparar distintas encuestas. Si viendo cómo está formulada la pregunta puedes hacerte una idea de lo que pretenden que respondas, la pregunta es capciosa; y las preguntas capciosas conducen a resultados sesgados. (En el capítulo 16 explico mejor cómo detectar problemas en encuestas.)

Buscar las mentiras en los lugares adecuados

De cuando en cuando te enteras de que alguien ha falseado datos o "manipulado cifras". Probablemente el engaño más frecuente relacionado con estadísticas y datos es cuando la gente descarta datos que no sustentan su hipótesis, no se ajustan al perfil o son muy diferentes al conjunto de información. En aquellos casos en que alguien ha cometido un error claro (por ejemplo, anotar que una persona tiene doscientos años), resulta apropiado eliminar el dato incorrecto o intentar corregir el error. Eliminar datos por cualquier otro motivo es éticamente inaceptable, pero ocurre a veces.

Cuando se comunican los resultados de un experimento, hay una frase recurrente: "De todas las personas que permanecieron hasta el final del estudio...". ¿Y qué pasa con las que abandonaron, sobre todo si es de medicina? ¿Se cansaron de los efectos secundarios del fármaco experimental y dejaron de tomarlo? En tal caso, la pérdida de esa persona supondrá un sesgo hacia un resultado positivo.

Antes de creerte los resultados de un estudio, entérate de cuántas personas fueron elegidas para participar, cuántas terminaron el estudio y qué ocurrió con todos los participantes, no sólo con los que experimentaron un resultado positivo.

Las encuestas, por su parte, tampoco no son inmunes a problemas causados por la falta de datos. Por ejemplo, los estadísticos saben que las opiniones de las personas que responden a una encuesta pueden ser muy diferentes de las opiniones de quienes no contestan. En general, cuanto menor es el porcentaje de personas que responden a una encuesta (la tasa de respuesta), menos fiables son los resultados. En el capítulo 16 encontrarás más información sobre encuestas y datos incompletos.

El impacto de las estadísticas engañosas

Todos los días, muchas veces sin darte cuenta siquiera, tomas decisiones basadas en estadísticas y estudios estadísticos de los que has oído hablar. Las estadísticas engañosas afectan a tu vida en mayor o menor medida según el tipo de estadísticas con las que tropieces y lo que decidas hacer con la información obtenida. Aquí tienes algunas situaciones cotidianas donde las estadísticas hacen acto de presencia:

- ✔ "Uf, espero que *Sultán* no muerda la alfombra otra vez mientras estoy en el trabajo. He leído en algún sitio que el Prozac ayuda a los animales de compañía a superar el trastorno de ansiedad por separación. ¿Cómo lo habrán descubierto? ¿Y qué dirían mis amigos si se enteraran de que medico a mi perro con un antidepresivo?".
- ✔ "Pensaba que todos teníamos que beber ocho vasos de agua al día, pero ahora me dicen que un exceso de agua puede ser perjudicial; ¿en qué quedamos?".
- ✔ "Un estudio dice que la gente pasa dos horas diarias en el trabajo leyendo y enviando correos electrónicos personales. ¿Cómo es posible? No me extraña que mi jefe esté paranoico".

También hay otras situaciones en que las estadísticas pueden tener una repercusión mayor en tu vida, y entonces es aún más importante ver las cosas de forma objetiva. Aquí tienes algunos ejemplos:

- ✔ Una plataforma que está presionando para que construyan un nuevo parque para monopatines te dice que el 80% de las personas encuestadas están de acuerdo en que les suban los impuestos para financiarlo, y que tú también deberías ser partidario. ¿Te sentirías presionado a decir que sí?
- ✔ En las noticias de la radio dicen que los teléfonos móviles causan cáncer cerebral. Tu mujer utiliza el móvil a todas horas. ¿Deberías alarmarte y tirar a la basura todos los móviles de tu casa?
- ✔ Ves un anuncio según el cual un determinado medicamento cura cierta enfermedad que tú padeces. ¿Vas corriendo a que el médico te extienda una receta?

Aunque no todas las estadísticas son engañosas y no todo el mundo quiere darte gato por liebre, debes estar atento. Si discriminas entre la información válida y la que es sospechosa o directamente rechazable puedes mantenerte alejado de las estadísticas equivocadas. Las herramientas y estrategias de este capítulo están pensadas para ayudarte a tomar distancia, analizar los problemas de forma crítica y tomar las decisiones correctas.

Capítulo 4

Herramientas del oficio

En este capítulo

- Ver la estadística como un proceso, no sólo como números
- Familiarizarse con los términos más importantes de la estadística

Hoy en día la palabra de moda es *datos*, como cuando uno dice: "¿Puedes aportar datos que respalden tu afirmación?", "¿Qué datos tenemos sobre esto?", "Los datos confirmaron la hipótesis original según la cual...", "Los datos estadísticos demuestran que..." o "Los datos me avalan". Pero la estadística es algo más que datos.

La estadística es todo el proceso de recopilación de pruebas para dar respuesta a preguntas sobre el mundo, en los casos en que esas pruebas resultan ser datos.

En este capítulo verás que la estadística funciona como un proceso y que los números tienen su propio papel. También te daré a conocer los términos técnicos más utilizados en estadística y te mostraré cómo esas definiciones y conceptos se enmarcan en este proceso. Así pues, la próxima vez que oigas decir que una encuesta tiene un margen de error de más/menos tres puntos porcentuales, tendrás al menos una noción básica de lo que eso significa.

Estadística: algo más que números

Los estadísticos no sólo "hacen estadísticas". Aunque el resto del mundo los considere unos tipos raros que trabajan con números, ellos se ven a sí mismos como los guardianes del método científico. Naturalmente, los estadísticos colaboran con expertos de otros campos para satisfacer su demanda de información, pero procesar los datos de otro es tan sólo una pequeña parte del trabajo de un estadístico (de hecho, si no hiciéramos otra cosa en todo el día dejaríamos nuestros empleos y nos haríamos asesores de los casinos). En realidad, la estadística interviene en todos los aspectos del *método científico*: formular preguntas idóneas, diseñar estudios, recopilar datos válidos, analizar esos datos correctamente y extraer conclusiones apropiadas. Pero, aparte de analizar los datos correctamente, ¿qué tienen que ver todos estos aspectos con la estadística? Eso es lo que explico en este capítulo.

Toda investigación comienza con una pregunta, por ejemplo:

> ✔ ¿Es perjudicial beber demasiada agua?
>
> ✔ ¿Cuál es el coste de la vida en San Francisco?
>
> ✔ ¿Quién ganará las próximas elecciones presidenciales?
>
> ✔ ¿Los productos de herbolario realmente ayudan a tener buena salud?
>
> ✔ ¿Renovarán mi serie de televisión favorita el año que viene?

Ninguna de estas preguntas habla directamente de cifras. Sin embargo, todas ellas requieren el uso de datos y procesos estadísticos para llegar a una respuesta.

Imagina que un investigador quiere saber quién ganará las próximas elecciones presidenciales en Estados Unidos. Para responder con conocimiento de causa, el investigador debe seguir varios pasos:

1. **Determinar la población objeto de estudio.**

   En este caso el investigador pretende estudiar a los ciudadanos mayores de edad que tengan previsto votar en las siguientes elecciones.

2. **Recopilar los datos.**

   Este paso es todo un desafío, porque no puedes salir a la calle y preguntar a todos y cada uno de los ciudadanos estadounidenses si tienen previsto votar y, en tal caso, por quién. Pero dejando eso de lado, pongamos que alguien dice: "Sí, tengo pensado ir a votar". ¿Sabes a ciencia cierta que esa persona acudirá a las urnas el día de las elecciones? ¿Y seguro que esa persona te dirá por quién va a votar? ¿Y qué pasa si esa persona cambia de opinión y vota por otro candidato?

3. **Organizar, resumir y analizar los datos.**

   Una vez el investigador ha recopilado todos los datos que necesita, organizarlos, resumirlos y analizarlos le ayuda a responder la pregunta

planteada. Este paso es el que la mayoría de la gente asocia con la estadística.

4. **Examinar todos los resúmenes de datos, gráficos, diagramas y análisis y extraer conclusiones de ellos para intentar responder la pregunta original.**

   Por supuesto, el investigador no estará seguro al 100% de que su respuesta sea correcta, ya que no habrá preguntado a todos y cada uno de los ciudadanos estadounidenses. No obstante, sí puede obtener una respuesta de la que esté seguro casi al 100%. De hecho, con una muestra de 2.500 personas seleccionadas de manera *no sesgada* (es decir, cuando todas las muestras posibles de 2.500 personas tienen la misma probabilidad de ser elegidas), el investigador puede obtener resultados exactos con un error de más/menos un 2,5% (si todos los pasos del proceso de investigación se llevan a cabo correctamente).

A la hora de extraer conclusiones, el investigador debe ser consciente de que todos los estudios tienen sus limitaciones y que, al existir siempre la posibilidad de cometer errores, los resultados podrían ser incorrectos. A este respecto, se puede proporcionar un valor numérico que indique a los demás cuánto confía el investigador en sus resultados y qué grado de exactitud se espera de ellos. (En el capítulo 12 encontrarás más información sobre el margen de error.)

Una vez ha concluido la investigación y se ha contestado la pregunta, es habitual que los resultados den pie a otras preguntas e investigaciones. Por ejemplo, si se observa que los hombres parecen favorecer a un candidato y las mujeres a otro, las siguientes preguntas podrían ser: "¿Quién responde a los sondeos más a menudo el día de las elecciones, los hombres o las mujeres? ¿Y qué factores determinan si acudirán a las urnas?".

En resumidas cuentas, la estadística consiste en aplicar el método científico para responder preguntas de investigación sobre el mundo. Los métodos estadísticos se utilizan en todos los pasos de un estudio, desde la fase de diseño hasta que se recopilan los datos, se organiza y resume la información, se lleva a cabo el análisis, se extraen conclusiones, se valoran las limitaciones y, por último, se diseña el siguiente estudio para responder las nuevas preguntas surgidas. La estadística es algo más que números: es un proceso.

Terminología básica de estadística

Todos los campos del saber tienen su propia jerga, y la estadística no es ninguna excepción. Si ves el proceso estadístico como una serie de etapas que atraviesas en el camino que va desde la pregunta hasta la respuesta, ya puedes suponer que en cada etapa te encontrarás con varias herramientas y términos técnicos (o jerga) con los que seguir adelante. Si se te están empezando a erizar los pelos de la nuca, tranquilízate. Nadie va a exigir que te conviertas en un experto en estadística ni que utilices esos términos a todas horas. Ni siquiera hace falta que vayas por ahí con una calculadora y un protector para el bolsillo de la camisa (los estadísticos no son así en realidad, es sólo una leyenda urbana). De todos modos, a medida que el mundo entero toma conciencia de la importancia de los números, los términos propios de la estadística se utilizan cada vez más en los medios de comunicación y en el lugar de trabajo, de manera que conocer su verdadero significado puede echarte un cable en muchas ocasiones. Además, si estás leyendo este libro porque quieres aprender a calcular algunas estadísticas, entender la terminología es el primer paso. Por eso en este apartado te doy a conocer algunos términos básicos del campo de la estadística. Para una explicación más detallada, sigue las referencias a los capítulos correspondientes de este libro.

Datos

Los *datos* son los elementos de información que recopilas durante el estudio. Por ejemplo, les pregunto a cinco amigos míos cuántas mascotas tienen y me dan los siguientes datos: 0, 2, 1, 4, 18 (el quinto amigo cuenta como mascotas todos los peces de su acuario). Pero no todos los datos son números; también tomo nota del sexo de todos mis amigos y obtengo lo siguiente: varón, varón, mujer, varón, mujer.

Existen dos grandes grupos de datos: los numéricos y los categóricos (a continuación comento a grandes rasgos las características de estas variables; en el capítulo 5 encontrarás una explicación más detallada).

- ✔ **Datos numéricos.** Estos datos tienen significado como medida, por ejemplo la altura, el peso, el coeficiente intelectual o la presión arterial; o bien son el resultado de un recuento, por ejemplo el número de acciones que posee una persona, los dientes de un perro o cuántas páginas de tu libro favorito eres capaz de leer antes de quedarte dormido (los estadísticos también los llaman *datos cuantitativos*).

  Los datos numéricos, a su vez, se dividen en dos tipos: discretos y continuos.

  - Los *datos discretos* representan elementos que pueden ser contados; adoptan valores posibles que se pueden enumerar. La lista de valores posibles puede estar restringida *(finita)* o puede ir desde 0, 1, 2 hasta el infinito *(infinita numerable)*. Por ejemplo, el número de caras obtenidas al lanzar 100 veces una moneda al aire adopta valores que van desde 0 hasta 100 (caso finito), pero el número de lanzamientos necesarios para sacar 100 caras adopta valores que van desde 100 (el número mínimo de lanzamientos) hasta el infinito. Los valores posibles son 100, 101, 102, 103... (representan el caso infinito numerable).

  - Los *datos continuos* representan mediciones; sus valores posibles no se pueden contar y tan sólo pueden describirse utilizando

intervalos de la recta de números reales. Por ejemplo, la cantidad exacta de gasolina que cargan los propietarios de vehículos con depósitos de 75 litros de capacidad puede adoptar cualquier valor posible entre 0,00 litros y 75,00 litros, lo cual se representa con el intervalo [0, 75] (bueno, en realidad sí que pueden contarse todos esos valores, pero ¿por qué ibas a hacerlo? En estos casos los estadísticos fuerzan un poquito la definición de continuo). Técnicamente, la vida útil de una pila AAA puede ser cualquier valor entre 0 e infinito. Por supuesto, no esperarás que dure más de unos pocos cientos de horas, pero nadie puede ponerle un límite concreto (¿te acuerdas del conejito de Duracell?).

- ✔ **Datos categóricos.** Los datos categóricos representan características como el sexo de una persona, su estado civil, su lugar de nacimiento o el tipo de películas que le gustan. Los datos categóricos pueden adoptar valores numéricos (por ejemplo el "1" para indicar un varón y el "2" para indicar una mujer), pero esos números no tienen un significado. No puedes sumarlos entre sí, por ejemplo (los datos categóricos también se denominan *datos cualitativos*).

Los *datos ordinales* combinan los datos numéricos y los categóricos. Los datos entran en dos categorías, pero los números asignados a esas categorías tienen significado. Por ejemplo, al valorar un restaurante en una escala de 0 a 5 se obtienen datos ordinales. Los datos ordinales generalmente se consideran categóricos, de manera que los grupos se ordenan al elaborar los gráficos y diagramas. Yo no me referiré a ellos de manera separada en este libro.

Conjunto de datos

Un *conjunto de datos* es la totalidad de los datos obtenidos de la muestra. Por ejemplo, si has pesado cinco envases y los pesos obtenidos son 6, 8, 11, 34 y 2 kilos, esos cinco números (6, 8, 11, 34, 2) forman tu conjunto de datos. Si únicamente anotas el tamaño del envase (por ejemplo pequeño, mediano o grande), el conjunto de datos podría ser el siguiente: mediano, mediano, mediano, grande, pequeño.

Variable

Una *variable* es una característica o un valor numérico que varía para cada individuo. Una variable puede representar el resultado de un recuento (por ejemplo, el número de mascotas que tienes) o una medición (el tiempo que tardas en levantarte por la mañana). O bien la variable puede ser categórica, de manera que cada persona se incluye en un grupo (o categoría) según unos criterios determinados (por ejemplo, filiación política, raza o estado civil). Los elementos de información registrados sobre unidades de análisis en relación con una variable son los datos.

Población

Para responder prácticamente a cualquier pregunta que se te ocurra investigar, debes enfocar tu atención a un grupo concreto de unidades de análisis (por ejemplo un grupo de personas, ciudades, animales, especímenes de roca, puntuaciones de examen, etc.). Por ejemplo:

- ✔ ¿Qué piensan los españoles de la política exterior de su gobierno?
- ✔ ¿Qué porcentaje de campos plantados fueron destrozados por ciervos el año pasado en el estado de Wisconsin?
- ✔ ¿Cuál es el pronóstico de las enfermas de cáncer de mama que toman un nuevo fármaco experimental?
- ✔ ¿Qué porcentaje de cajas de cereales llevan la cantidad de producto que consta en el envase?

En cada uno de estos ejemplos se plantea una pregunta. Y en cada caso puedes identificar un grupo concreto de unidades de análisis: los ciudadanos españoles, todos los campos plantados de Wisconsin, todas las enfermas de cáncer de mama y todas las cajas de cereales, respectivamente. El grupo de elementos o unidades que quieres estudiar para responder a la pregunta que da pie a la investigación es lo que denominamos *población*. Sin embargo, a veces cuesta mucho definir una población. En un buen estudio, los investigadores definen la población de forma muy clara, mientras que en uno malo la población no está bien definida.

La pregunta de si los bebés duermen mejor con música es un buen ejemplo de lo difícil que puede ser definir la población. ¿Cómo definirías exactamente a un bebé? ¿Un niño de menos de tres meses? ¿De menos de un año? ¿Y quieres estudiar sólo a los bebes de un determinado país o a los bebés de todo el mundo? Los resultados pueden variar entre bebés de distintas edades, entre bebés españoles y japoneses, etc.

Muchas veces los investigadores quieren estudiar una población muy grande, pero al final (para ahorrar tiempo o dinero o porque no saben hacerlo mejor) acaban estudiando una muy pequeña. Este atajo puede suponer un gran problema a la hora de extraer conclusiones. Por ejemplo, pongamos que un profesor de universidad quiere estudiar el modo en que los anuncios de televisión persuaden a los consumidores para que compren productos. Su estudio está basado en un grupo de sus propios alumnos que participaron para que les subiera medio punto la nota final de la asignatura. Este grupo experimental puede ser adecuado, pero los resultados no pueden extrapolarse a una población distinta de sus alumnos, ya que en el estudio no está representada más que la reducida población del aula.

Muestras y aleatoriedad

¿Qué haces cuando pruebas un cocido? Remueves la olla con la cuchara, coges un poquito y lo pruebas. A continuación extraes una conclusión sobre el contenido de la olla entera, aunque no lo hayas probado todo. Si tomas la muestra de forma imparcial (por ejemplo, no te limitas a probar sólo los ingredientes más sabrosos), puedes hacerte una idea clara del sabor del cocido sin tener que comértelo todo. En estadística las muestras se toman exactamente igual. Los investigadores quieren averiguar algo sobre una población pero no tienen suficiente tiempo o dinero para estudiar a todos los elementos de esa población, de manera que eligen un subconjunto de elementos, los estudian y utilizan la información obtenida para extraer conclusiones sobre toda la población. Este subconjunto de la población es lo que se denomina *muestra*.

Aunque seleccionar una muestra parece algo muy sencillo, en realidad no lo es. La manera de seleccionar una muestra de la población puede marcar la diferencia entre obtener resultados correctos y acabar con un montón de basura. Un ejemplo: pongamos que quieres una muestra de lo que opinan los adolescentes sobre el tiempo que pasan en Internet. Si envías una encuesta con un mensaje de texto, los resultados no representarán las opiniones de todos los adolescentes, que es la población que quieres estudiar. Únicamente representarán a los adolescentes que tengan acceso a mensajes de texto. ¿Ocurre a menudo este desajuste estadístico? No te quepa duda.

Las encuestas a través de Internet son un claro ejemplo de tergiversación estadística por culpa de una mala selección de la muestra. En la red hay miles de encuestas de opinión en las que se puede participar visitando un determinado sitio web. Pero incluso si contestaran una encuesta en Internet 50.000 españoles, esa muestra no sería representativa de toda la población de España; tan sólo representaría a las personas que tuvieran acceso a Internet, visitaran ese sitio web en particular y estuvieran suficientemente interesadas en el tema como para participar en la encuesta (lo cual generalmente significa que tienen opiniones tajantes al respecto). El resultado de todos estos problemas es el *sesgo*, el favoritismo sistemático de determinadas personas o determinados resultados del estudio.

¿Cómo debe seleccionarse una muestra de forma que se evite el sesgo? La palabra clave es *aleatoriedad*. Una *muestra aleatoria* es aquella que se selecciona con igualdad de oportunidades, es decir, cada muestra posible del mismo tamaño que la tuya tiene la misma probabilidad de ser seleccionada de la población. El significado real de *aleatorio* es que ningún grupo de población se ve favorecido ni excluido del proceso de selección.

Las muestras *no aleatorias* (o sea, mal hechas) son las que se seleccionan con algún tipo de preferencia o de exclusión automática de una parte de la población. Un ejemplo típico de muestra no aleatoria es la generada por las encuestas de llamada voluntaria *(call-in)*, en las que es la gente quien llama por teléfono y da su opinión sobre un tema en particular en respuesta a un llamamiento de una cadena de televisión. Las personas que deciden participar en este tipo de encuestas no representan a la población en general porque necesariamente estaban viendo el programa y, además, el tema les interesaba lo suficiente como para llamar. Técnicamente no representan una muestra, en el sentido estadístico de la palabra, porque nadie las ha seleccionado, sino que se han elegido a sí mismas para participar, creando así una *muestra de respuesta voluntaria* o *muestra autoseleccionada*. Los resultados presentarán un sesgo hacia las personas que tengan opiniones tajantes sobre esa cuestión.

Para tomar una muestra auténticamente aleatoria necesitas un mecanismo de aleatorización para elegir a los participantes. Por ejemplo, la organización Gallup empieza con una lista informatizada de todas las centralitas telefónicas de Estados Unidos, junto con estimaciones del número de viviendas que tienen conexión telefónica. El ordenador utiliza un procedimiento llamado *marcación aleatoria de dígitos* (RDD por sus siglas en inglés) para generar al azar una relación exhaustiva de los números de teléfono de las centralitas, y a continuación selecciona muestras de esos números. Es decir, el ordenador crea una lista de todos los números de teléfono posibles de Estados Unidos y luego selecciona un subconjunto de números de esa lista para que Gallup llame.

Otro ejemplo de muestreo aleatorio es el empleo de generadores de números aleatorios. En este proceso, los elementos de la muestra se eligen utilizando una lista de números aleatorios generados por ordenador; de manera que cada muestra de elementos tiene la misma probabilidad de ser seleccionada. Los investigadores pueden utilizar este método de aleatorización para asignar enfermos al grupo experimental y al grupo de control en un experimento. Este proceso es el equivalente a extraer nombres de un sombrero o extraer los números del bombo de la lotería.

Por muy grande que sea una muestra, si está basada en métodos no aleatorios los resultados no serán representativos de la población sobre la cual el investigador quiere extraer conclusiones. No te dejes impresionar por el tamaño de una muestra: primero comprueba cómo la han seleccionado. Busca el término *muestra aleatoria*. Si ese término aparece, lee la letra pequeña para saber exactamente qué método utilizaron y aplica la definición anterior para cerciorarte de que esa muestra realmente se eligió de forma aleatoria. Una muestra aleatoria pequeña es mejor que una muestra no aleatoria grande.

Estadístico

Un *estadístico* es un número que resume los datos recopilados de una muestra. Existen muchos estadísticos diferentes para resumir datos. Por

ejemplo, los datos se pueden resumir como porcentaje (el 60% de las unidades familiares de la muestra tienen más de dos coches), promedio (el precio medio de una vivienda en esta muestra es de...), mediana (el sueldo mediano de los 1.000 ingenieros informáticos que componen la muestra es de...) o percentil (este mes tu bebé está en el percentil 90 de peso según datos recopilados de más de 10.000 bebés).

El tipo de estadístico que se calcule depende del tipo de datos. Por ejemplo, los porcentajes se utilizan para resumir datos categóricos y las medias se utilizan para resumir datos numéricos. El precio de una casa es una variable numérica, de manera que puedes calcular su media o su desviación estándar. Por el contrario, el color de una casa es una variable categórica; no tiene sentido buscar la desviación estándar ni la media de color. En este caso los estadísticos importantes son los porcentajes de casas de cada color.

No todas las estadísticas son correctas ni ecuánimes, por supuesto. Sólo porque alguien te dé una estadística, no tienes la garantía de que esa estadística sea científica o legítima. A lo mejor has oído alguna vez el siguiente dicho: "Los números no mienten, pero los mentirosos también usan números".

Parámetro

Los estadísticos se basan en datos de una muestra, no en datos de la población. Si recopilas datos de la población entera, estás haciendo un *censo*. Si a continuación resumes toda la información del censo en un único número procedente de una variable, ese número es un *parámetro*, no un estadístico. Las más de las veces, los investigadores intentan estimar los parámetros utilizando estadísticos. El Instituto Nacional de Estadística español quiere dar a conocer el número total de habitantes del país, así que lleva a cabo un censo. Sin embargo, debido a los problemas logísticos que entraña una tarea tan ardua (por ejemplo ponerse en contacto con las personas sin hogar), a la postre las cifras del censo sólo pueden llamarse *estimaciones*, y se corrigen al alza para dar cuenta de las personas no incluidas en el censo.

Sesgo

La palabra *sesgo* se utiliza bastante a menudo, y probablemente ya sepas que significa algo malo. Pero ¿qué es el sesgo exactamente? El *sesgo* es una inclinación sistemática que está presente en el proceso de recopilación de datos y que da lugar a resultados desviados y engañosos. El sesgo puede aparecer por varias razones:

- ✔ **Por la manera de seleccionar la muestra.** Por ejemplo, si quieres obtener una estimación del dinero que tienen previsto gastar los ciudadanos estadounidenses en las compras de Navidad, y para ello coges tu carpeta, te diriges a un centro comercial el día siguiente al Día de Acción de Gracias y le preguntas a la gente cuáles son sus planes de compra, está claro que el proceso de muestreo estará sesgado. La muestra tenderá a favorecer a los buscadores de gangas que acuden a ese centro comercial en particular a luchar a brazo partido con cientos de otros clientes el día del año que en Estados Unidos se conoce como "viernes negro".
- ✔ **Por la manera de recopilar los datos.** Las preguntas de las encuestas son una causa importante de sesgo. Como muchas veces los investigadores buscan un resultado en concreto, las preguntas que hacen pueden reflejar y conducir a ese resultado esperado. Por ejemplo, la cuestión de si procede financiar con dinero público los colegios concertados siempre genera un gran debate, con vehementes opiniones tanto a favor como en contra. La pregunta "¿No cree que sería una buena inversión de futuro ayudar a los colegios concertados?" tiene un poco de sesgo. Y lo mismo ocurre con la pregunta "¿No está usted cansado de pagar dinero de su bolsillo para educar a los hijos de otros?". La manera de formular la pregunta puede influir mucho en los resultados.

Otras cosas que provocan sesgo son el momento de hacer la encuesta, la longitud y la dificultad de las preguntas, y la forma de ponerse en contacto con las personas de la muestra (por teléfono, por correo postal, de puerta en puerta, etc.). En el capítulo 16 encontrarás más información sobre la manera correcta de diseñar y evaluar encuestas y sondeos.

Cuando analices los resultados de una encuesta que sea importante para ti o en la que tengas un interés especial, antes de extraer conclusiones sobre los resultados entérate de qué preguntas plantearon y cómo las formularon exactamente.

Media (Promedio)

La media, también llamada *promedio*, es el estadístico más utilizado para medir el centro de un conjunto de datos numérico. La *media* es la suma de todos los números dividida por la cantidad total de números. La media de la población entera se denomina *media poblacional*, y la media de una muestra se denomina, lógicamente, *media muestral*. (En el capítulo 5 hablo de la media con más detalle.)

La media puede no ser una representación ecuánime de los datos, ya que se ve influenciada fácilmente por los *valores atípicos* (valores muy

grandes o muy pequeños que se alejan mucho de otros valores del conjunto de datos).

Mediana

La mediana es otra manera de medir el centro de un conjunto de datos numéricos. La mediana estadística viene a ser como la mediana de una autopista. En la mayoría de las autopistas la mediana es el centro, de manera que en ambos lados de ella hay el mismo número de carriles. En un conjunto de datos numéricos, la *mediana* es el valor que divide dicho conjunto en dos partes iguales, una con valores superiores y otra con valores inferiores al valor de la mediana. Así pues, la mediana es el centro auténtico del conjunto de datos. En el capítulo 5 encontrarás más información sobre la mediana.

La próxima vez que te den la cifra correspondiente a una media, mira a ver si también te dicen cuál es la mediana. Si no es así, ¡pide que te la den! La media y la mediana son dos representaciones distintas del centro de un conjunto de datos y a menudo cuentan historias muy diferentes sobre los datos, sobre todo cuando el conjunto de datos contiene valores atípicos (cifras muy grandes o muy pequeñas que se salen de lo normal).

Desviación estándar (o típica)

¿Alguna vez has oído a alguien decir que un resultado en particular presenta "dos desviaciones estándares por encima de la media"? Cada vez más, la gente quiere comunicar la importancia de los resultados que han obtenido, y una manera de hacerlo es diciendo el número de desviaciones estándares por encima o por debajo de la media. Pero ¿qué es exactamente una desviación estándar?

La *desviación estándar* es una medida utilizada por los estadísticos para referirse al grado de variabilidad (o dispersión) de los números de un conjunto de datos. Como el propio término indica, la desviación estándar es el grado estándar (o típico) de desviación (o divergencia) respecto del promedio (o la media, como prefieren llamarla los estadísticos). Así pues, dicho de manera muy burda, la desviación estándar es la divergencia media respecto de la media.

La fórmula de la desviación estándar (representada por la letra *s*) es la siguiente, donde *n* es el número de valores del conjunto de datos, cada una de las *x* representa a un número del conjunto de datos, y $\bar{x}$ es la media de todos los datos:

$$s = \sqrt{\sum \frac{(x-\bar{x})^2}{n-1}}$$

En el capítulo 5 encontrarás instrucciones detalladas para calcular la desviación estándar.

La desviación estándar también se utiliza para describir la zona donde deberían estar la mayoría de los datos, en relación con la media. Por ejemplo, si los datos siguen una curva con forma de campana (lo que se llama *distribución normal*), aproximadamente el 95% de los valores se encuentran a no más de dos desviaciones estándares de la media (este resultado es lo que se llama *regla empírica*, o *regla 68-95-99,7*; tienes más información sobre ella en el capítulo 5).

La desviación estándar es un estadístico importante, pero a menudo se omite al comunicar resultados estadísticos. Si no te la dan, te están contando sólo una parte de la verdad. A los estadísticos les gusta contar la anécdota de un hombre que tenía un pie metido en un cubo de agua helada y el otro pie en un cubo de agua hirviendo. En promedio estaba la mar de a gusto, pero piensa en la variabilidad de las dos temperaturas de los pies. Si buscamos ejemplos más próximos, el precio medio de una vivienda no te dice nada sobre el abanico de precios que puedes encontrar en el mercado, y el sueldo medio quizá no represente del todo la situación real de tu empresa en el supuesto de que los sueldos sean muy dispares.

No te conformes con saber la media. Asegúrate de preguntar también cuál es la desviación estándar. Sin la desviación estándar no tienes manera de conocer el grado de dispersión de los valores (si estás hablando de sueldos iniciales, por ejemplo, esto podría ser muy importante).

Percentil

Probablemente hayas oído hablar antes de los percentiles. Si alguna vez has hecho alguna prueba normalizada, junto con la puntuación obtenida

debieron de darte una medida de tu resultado en comparación con el resto de las personas que hicieron la prueba. Esta medida comparativa probablemente te fue comunicada en forma de percentil. El *percentil* para un dato concreto es el porcentaje de valores de la muestra que están por debajo de ese dato concreto. Por ejemplo, si te dicen que tu puntuación está en el percentil 90, significa que el 90% de las personas que hicieron la misma prueba obtuvieron una puntuación inferior a la tuya (y el 10% de los que se examinaron obtuvieron una puntuación más alta que tú). La mediana está justo en el centro de un conjunto de datos, de manera que representa el percentil 50. En el capítulo 5 encontrarás más información sobre percentiles.

Los percentiles se utilizan de varias formas con fines de comparación y para determinar la *posición relativa* (es decir, la situación de un valor en concreto en comparación con el resto de los valores). El peso de los bebés suele indicarse como percentil, por ejemplo. Las empresas también utilizan los percentiles para saber en qué situación se encuentran, en comparación con otras empresas, en cuanto a ventas, beneficios, satisfacción del cliente, etc.

Puntuación estándar (o típica)

La puntuación estándar es una manera hábil de poner en perspectiva unos resultados sin tener que dar un montón de detalles (algo que a los medios de comunicación les encanta hacer). La *puntuación estándar* representa el número de desviaciones estándares por encima o por debajo de la media (sin preocuparse de cuál es la desviación estándar ni la media).

Por ejemplo, imagina que Roberto ha obtenido una puntuación de 400 en una prueba de acceso a la universidad. ¿Qué significa eso? Pues no mucho, porque no puedes poner ese 400 en perspectiva. Pero si te dicen que la puntuación estándar de Roberto en la prueba es +2, ya lo sabes todo. Sabes que su puntuación está dos desviaciones estándares por encima de la media (¡bien hecho, Roberto!). Ahora imagina que la puntuación estándar de Isabel es –2. En este caso el resultado no es bueno (para Isabel), ya que su nota está dos desviaciones estándares por debajo de la media.

El proceso de convertir un número en una puntuación estándar se llama *normalización o estandarización*. En el capítulo 9 explico la manera de calcular e interpretar puntuaciones estándares cuando la distribución es normal (con forma de campana).

Distribución y distribución normal

La *distribución* de un conjunto de datos (o de una población) es una lista o función que muestra todos los valores posibles (o intervalos) de los datos y la frecuencia con la que aparecen. Cuando se organiza una distribución de datos categóricos, ves el número o porcentaje de elementos que hay en cada grupo. Cuando se organiza una distribución de datos numéricos, generalmente se ordenan del más pequeño al más grande, se dividen en grupos de tamaño razonable (si conviene) y luego se pasan a gráficos y diagramas para ver la forma, el centro y el grado de variabilidad de los datos.

Existen muchas distribuciones distintas para datos categóricos y numéricos, y las más comunes tienen sus propios nombres. Una de las distribuciones más conocidas es la *distribución normal*, cuya representación gráfica es la popular *campana de Gauss* (también llamada curva gaussiana o curva de campana). La distribución normal se basa en datos numéricos continuos, cuyos valores posibles se encuentran en la recta numérica real. Cuando los datos se organizan a manera de gráfico, esta distribución tiene forma de campana simétrica. Dicho de otro modo, la mayoría (el 68%) de los datos están centrados en torno a la media (la parte central de la campana), y a medida que te alejas de la media hacia los lados encuentras cada vez menos valores (las curvas descendentes de ambos lados de la campana).

La media (y, por tanto, la mediana) está justo en el centro de la distribución normal debido a la simetría, y la desviación estándar se mide por la distancia desde la media hasta el *punto de inflexión* (el punto donde la curva cambia de ser convexa a ser cóncava). La figura 4-1 muestra un gráfico de una distribución normal en la que la media es 0 y la desviación estándar es 1 (esta distribución tiene un nombre especial, la *distribución normal estándar* o *distribución Z*). La curva tiene forma acampanada.

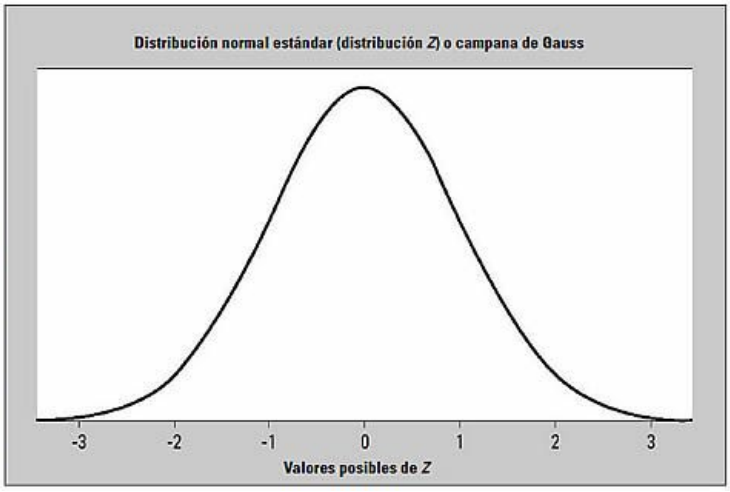


Figura 4-1: Una distribución normal estándar (distribución *Z*) tiene forma acampanada con media 0 y desviación estándar 1

Puesto que cada población de datos tiene su propia media y su propia desviación estándar, existe un número infinito de distribuciones normales, cada una con una media y una desviación estándar que la caracterizan. En el capítulo 9 encontrarás muchísima más información sobre la

distribución normal y la normal estándar.

Teorema del límite central

La distribución normal también se utiliza para medir la exactitud de muchos estadísticos, incluida la media, por medio de un importante resultado llamado *teorema del límite central*. Este teorema permite medir cuánto variará la media muestral sin tener que compararla con otras medias muestrales (¡menos mal!). Teniendo en cuenta esta variabilidad podemos utilizar los datos para responder a preguntas sobre la población, por ejemplo, "¿Cuáles son los ingresos medios por unidad familiar en España?" o "este informe dice que el 75% de las tarjetas regalo no llegan a utilizarse nunca, ¿es eso cierto?". (Estos dos análisis en particular, posibles gracias al teorema del límite central, se llaman *intervalos de confianza* y *contrastes de hipótesis*, respectivamente, y se describen en los capítulos 13 y 14, respectivamente.)

El teorema del límite central (abreviado *TLC*) dice básicamente que, para datos no normales, la media muestral presenta una distribución aproximadamente normal con independencia de cómo sea la distribución de los datos originales (siempre que la muestra sea suficientemente grande). Y esto no se aplica sólo a la media muestral; el TLC también se cumple para otros estadísticos como, por ejemplo, la proporción muestral (ver capítulos 13 y 14). Como los estadísticos conocen a la perfección la distribución normal (tratada en el apartado anterior), estos análisis son mucho más sencillos. En el capítulo 11 explico mejor el teorema del límite central, conocido en el mundillo como "la joya de la Corona en el campo de la estadística" (estos estadísticos son unos tíos raros de verdad).

Valores z

Si un conjunto de datos tiene una distribución normal y tú normalizas todos los datos para obtener puntuaciones estándares, esas puntuaciones estándares se llaman valores *z*. Todos los valores *z* presentan lo que se conoce como distribución normal estándar (o distribución *Z*). La *distribución normal estándar* es una distribución normal especial donde la media es igual a 0 y la desviación estándar es igual a 1.

La distribución normal estándar resulta útil para analizar los datos y determinar estadísticos como percentiles, o el porcentaje de datos que se encuentra entre dos valores. Así pues, si los investigadores determinan que los datos poseen una distribución normal, generalmente empiezan por normalizar los datos (convirtiendo cada punto de datos en un valor *z*) y luego utilizan la distribución normal estándar para analizar los datos más a fondo. En el capítulo 9 encontrarás más información sobre los valores *z*.

Experimentos

Un *experimento* es un estudio que impone un tratamiento (o control) a los sujetos (participantes), controla su entorno (por ejemplo limitando su alimentación, administrándoles cierta dosis de un fármaco o placebo o pidiéndoles que permanezcan despiertos durante un tiempo determinado) y registra las respuestas. El propósito de la mayoría de los experimentos es encontrar una relación causa-efecto entre dos factores (por ejemplo el consumo de alcohol y la vista defectuosa, o la dosis de un fármaco y la intensidad de sus efectos secundarios). Aquí tienes algunas preguntas típicas que los experimentos intentan responder:

- ✔ ¿Tomar zinc ayuda a reducir la duración de un resfriado? Algunos estudios dicen que sí.
- ✔ ¿La forma y la posición de la almohada influyen en el descanso nocturno? Los especialistas del Centro Emory sobre investigación de la columna vertebral, en Atlanta, dicen que sí.
- ✔ ¿La altura de los tacones afecta a la comodidad de los pies? Un estudio de la universidad de UCLA dice que es mejor usar tacones de dos centímetros que zapatos totalmente planos.

En este apartado comento otras varias definiciones de palabras que quizá oigas cuando alguien hable sobre experimentos. El capítulo 17 trata exclusivamente sobre esta cuestión. Por ahora, céntrate sólo en la terminología básica de los experimentos.

Grupo experimental y grupo de control

La mayoría de los experimentos intentan dilucidar si un determinado tratamiento experimental (o factor importante) tiene un efecto significativo en un resultado. Por ejemplo, ¿el zinc ayuda a reducir la duración de un resfriado? Los sujetos que participan en el experimento suelen dividirse en dos grupos: un grupo experimental y un grupo de control (también puede haber más de un grupo experimental).

- ✔ El *grupo experimental* está compuesto por personas que toman el tratamiento experimental cuyos efectos se quiere estudiar (en este caso, comprimidos de zinc).
- ✔ El *grupo de control* está compuesto por personas que no tomarán los comprimidos de zinc. En su lugar se les administra un placebo

> (un tratamiento ficticio, por ejemplo una pastilla de azúcar), un tratamiento no experimental (por ejemplo vitamina C, en el estudio sobre el zinc) o nada en absoluto, según la situación.

Al final, las respuestas de los integrantes del grupo experimental se comparan con las respuestas del grupo de control para buscar diferencias estadísticamente significativas (diferencias que difícilmente puedan deberse al azar).

Placebo

Un *placebo* es un tratamiento ficticio, por ejemplo una pastilla de azúcar. Los placebos se administran al grupo de control para producir un fenómeno psicológico llamado *efecto placebo*, que consiste en que algunas personas presentan una respuesta como si se tratara de un tratamiento real. Por ejemplo, después de tomar una pastilla de azúcar, una persona que experimente el efecto placebo puede decir: "Sí, ya me siento mejor", o "Vaya, estoy empezando a marearme un poco". Midiendo el efecto placebo en el grupo de control puedes averiguar qué informes del grupo experimental son reales y cuáles se deben probablemente al efecto placebo (los experimentadores dan por sentado que el efecto placebo afecta tanto al grupo experimental como al grupo de control).

Enmascarado y doblemente enmascarado

Un *experimento enmascarado* (a veces se llama "a ciegas" o "ciego") es un estudio clínico en el que los sujetos participantes no saben si están en el grupo experimental (el que recibe el tratamiento) o en el de control. Siguiendo con el ejemplo del zinc, los investigadores se asegurarían de que los comprimidos de vitamina C y los comprimidos de zinc tuvieran exactamente el mismo aspecto, y no dirían a los pacientes cuál de los dos tipos les estarían administrando. Un experimento enmascarado intenta controlar el sesgo por parte de los participantes.

Un *experimento doblemente enmascarado*, o *con doble enmascaramiento*, controla el sesgo potencial por parte de los pacientes y de los investigadores. Ni los pacientes ni los investigadores que recopilan los datos saben qué sujetos han recibido el tratamiento y cuáles no. Entonces, ¿quién está enterado? Generalmente es un tercero (alguien que no participa en el experimento de ninguna otra forma) quien se encarga de juntar las piezas. Un estudio con doble enmascaramiento es mejor porque, aunque los investigadores aseguren actuar de forma no sesgada, a menudo tienen un interés especial en los resultados (¡de lo contrario no estarían haciendo el estudio!).

Encuestas (sondeos)

Una *encuesta* (a veces llamada *sondeo*) es un cuestionario; generalmente se utiliza para obtener las opiniones de la gente junto con algunos datos demográficos relevantes. Al haber tantos responsables políticos, expertos en marketing y otras personas que quieren "tomar el pulso a la población" y averiguar lo que el ciudadano medio piensa y siente, mucha gente tiene la sensación de no poder escapar a esa persecución. Probablemente tú mismo hayas recibido muchas solicitudes de participación en encuestas, y puede que hasta te hayas vuelto inmune a ellas y te limites a tirar a la basura los cuestionarios que te llegan por correo o a negarte cuando te piden que participes en una encuesta telefónica.

Si se hace como es debido, una encuesta puede ser muy informativa. Las encuestas se utilizan para averiguar qué programas de televisión gustan más, qué piensan los consumidores sobre las compras por Internet y si en Estados Unidos deberían permitir que alguien menor de treinta y cinco años llegue a ser presidente, por ejemplo. Las empresas se sirven de las encuestas para valorar el grado de satisfacción de sus clientes, para averiguar qué productos quiere la gente y para determinar quién compra sus artículos. Las cadenas de televisión utilizan las encuestas para conocer las reacciones inmediatas a noticias que acaban de salir en antena, y los productores cinematográficos las usan para decidir cómo debería terminar una película.

Sin embargo, si tuviera que elegir una palabra para valorar el estado general de las encuestas en los medios de comunicación, creo que sería "cantidad", y no "calidad". Dicho de otro modo: hay montones de encuestas mal hechas. Por suerte para ti, en este libro encontrarás montones de buenos consejos e información para analizar, criticar y comprender los resultados de las encuestas, y también para diseñar tus propias encuestas y hacer las cosas bien. (Si quieres meterte ya en materia, pasa al capítulo 16.)

Margen de error

Posiblemente hayas visto u oído resultados como el siguiente: "Esta encuesta tiene un margen de error de más/menos tres puntos porcentuales". ¿Qué significa eso? La mayoría de las encuestas (salvo los censos) se basan en información recopilada de una muestra de personas, no de la población entera. Forzosamente existirá cierto grado de error, y no me refiero a un error de cálculo (aunque también puede haberlo), sino a un *error de muestreo* (también llamado *error muestral*), que ocurre simplemente porque los investigadores no están preguntando a todo el mundo. El *margen de error* mide la diferencia máxima que puede haber entre los resultados de la muestra y los resultados de la población real. Puesto que los resultados de la mayoría de las preguntas pueden expresarse como porcentajes, el margen de error casi siempre se indica también como porcentaje. ¿Cómo se interpreta un margen de error? Pongamos que sabes que el 51% de las personas de la muestra han dicho que piensan votar por la señora Cálculo en las próximas elecciones. Si quisieras extrapolar esos resultados a todos los votantes, tendrías que sumar y restar el margen de error y proporcionar un intervalo de resultados posibles para estar suficientemente seguro de que salvas la distancia existente entre la muestra y la población entera. Suponiendo que el margen de error es de más/menos tres puntos porcentuales, estarías bastante seguro de que entre el 48% (51–3) y el 54% (51+3) de la población votará por la señora Cálculo en las elecciones, basándote en los resultados muestrales. En este caso la señora Cálculo podría obtener un poco más o un poco menos de la mayoría de los votos, de manera que podría ganar o perder las elecciones. Esta situación se ha repetido en varias noches electorales, cuando los medios de comunicación han querido adelantar quién era el ganador pero, según las encuestas a pie de urna, el resultado estaba "muy reñido". En el capítulo 12 encontrarás más información sobre el margen de error.

El margen de error mide la exactitud; no mide la cantidad de sesgo que pueda haber (me he referido al sesgo anteriormente en este mismo capítulo). Unos resultados que sean numéricamente exactos no significan nada en absoluto si se han recopilado de forma sesgada.

Intervalo de confianza

Una de las aplicaciones más importantes de la estadística consiste en estimar un parámetro poblacional utilizando un valor muestral. O dicho de otro modo: utilizar un número que resume una muestra para ayudarte a estimar el número correspondiente que resume a toda la población (anteriormente en este mismo capítulo he puesto las definiciones de parámetro y estadístico). En cada una de las siguientes preguntas estás buscando un parámetro poblacional:

- ✔ ¿Cuáles son los ingresos medios por unidad familiar en Azerbaiyán? (Población=todas las unidades familiares de Azerbaiyán; parámetro=ingresos medios por unidad familiar.)
- ✔ ¿Qué porcentaje de europeos vieron este año la ceremonia de entrega de los Oscar? (Población=todos los europeos; parámetro=porcentaje que vieron este año la ceremonia de entrega de los Oscar.)
- ✔ ¿Cuál es la esperanza media de vida de un bebé que nazca hoy? (Población=todos los bebés que nazcan hoy; parámetro=esperanza media de vida.)
- ✔ ¿Qué eficacia tiene este nuevo medicamento para los adultos con alzhéimer? (Población=todas las personas que padezcan alzhéimer; parámetro=porcentaje de esas personas que experimenten alguna mejoría al tomar ese medicamento.)

Es imposible conocer esos parámetros con exactitud; cada uno de ellos requiere una estimación basada en una muestra. Primero se toma una muestra aleatoria de una población (pongamos una muestra de 1.000 unidades familiares de Azerbaiyán) y a continuación se encuentra el estadístico muestral correspondiente (los ingresos medios por unidad familiar de la muestra). Como sabes que los resultados varían para cada muestra, tienes que añadir un "más/menos algo" a los resultados de la muestra si quieres extraer conclusiones sobre toda la población (todas las unidades familiares de Azerbaiyán). Este "más/menos" que añades al estadístico muestral para estimar un parámetro es el margen de error.

Cuando coges un estadístico de la muestra (por ejemplo la media muestral o un porcentaje muestral) y sumas/restas un margen de error, obtienes lo que en estadística se llama *intervalo de confianza*. Un intervalo de confianza representa un intervalo de valores probables para el parámetro poblacional, a partir del estadístico muestral. Por ejemplo, pongamos que todos los días tardas una media de treinta y cinco minutos en ir de casa al trabajo, con un margen de error de más/menos cinco minutos. Puedes estimar que el tiempo medio que tardas en llegar al trabajo está comprendido entre treinta y cuarenta minutos. Esta estimación es un intervalo de confianza.

Algunos intervalos de confianza son mayores que otros (y cuanto mayor sea, peor, porque la precisión será menor). Existen varios factores que influyen en la amplitud de un intervalo de confianza, por ejemplo el tamaño de la muestra, el grado de variabilidad de la población estudiada y la confianza que quieres tener en los resultados (la mayoría de los investigadores se contentan con tener un 95% de confianza en sus resultados). En el capítulo 13 encontrarás más factores que influyen en los intervalos de confianza, así como instrucciones para calcular e interpretar intervalos de confianza.

Contrastes de hipótesis

Es probable que en tu devenir diario con números y estadísticas no te hayas tropezado nunca con el término *contraste de hipótesis*. Sin embargo, puedo asegurarte que los contrastes de hipótesis tienen una gran influencia en tu vida personal y profesional, simplemente por el papel tan importante que desempeñan en la industria, la medicina, la agricultura, el gobierno y muchos otros ámbitos. Cada vez que oyes a alguien decir que su estudio presenta un resultado "estadísticamente significativo" te tropiezas con un contraste de hipótesis (un resultado estadísticamente significativo es uno que difícilmente puede haber ocurrido por casualidad; tienes más información sobre este asunto en el capítulo 14).

Básicamente, un *contraste de hipótesis* es un procedimiento estadístico mediante el cual se recopilan datos de una muestra y se cotejan con una afirmación referida a un parámetro poblacional. Por ejemplo, si una cadena de pizzerías asegura que entrega todas las pizzas en un tiempo máximo de treinta minutos tras recibir el pedido, como media, podrías comprobar si esa afirmación es cierta recopilando una muestra de tiempos de entrega durante un determinado período y determinando el tiempo medio de entrega para esa muestra. Para tomar una decisión también debes tener en cuenta cuánto pueden variar tus resultados de una muestra a otra (lo cual está relacionado con el margen de error).

Puesto que tu decisión se basa en una muestra y no en la población entera, el contraste de hipótesis puede conducirte a veces a una conclusión errónea. Sin embargo, la estadística es todo lo que tienes, y si la utilizas en la forma debida tendrás muchas posibilidades de acertar. En el

capítulo 14 encontrarás más información sobre contrastes de hipótesis.

En los estudios científicos se realizan muchos contrastes de hipótesis, incluidas pruebas *t* (comparan dos medias poblacionales), pruebas *t* para datos apareados (se examinan las diferencias entre el antes y el después) y pruebas de afirmaciones referidas a proporciones o medias de una o más poblaciones. En el capítulo 15 encontrarás información más concreta sobre estos contrastes de hipótesis.

Valores *p*

Los contrastes de hipótesis sirven para verificar la validez de una afirmación referida a una población. Esa afirmación que se somete a juicio se llama *hipótesis nula*. La *hipótesis alternativa* es la que creerías si concluyeras que la hipótesis nula está equivocada. Las pruebas de este juicio son los datos y los estadísticos que los acompañan. Todos los contrastes de hipótesis utilizan un valor *p* para ponderar la solidez de las pruebas (lo que los datos te están diciendo sobre la población). El valor *p* es un número comprendido entre 0 y 1 que se interpreta de la manera siguiente:

- ✔ Un valor *p* pequeño (por lo general, ≤0,05) indica una prueba sólida en contra de la hipótesis nula, de manera que puedes rechazar dicha hipótesis.
- ✔ Un valor *p* grande (por lo general,>0,05) indica una prueba débil en contra de la hipótesis nula, de manera que no rechazas dicha hipótesis.
- ✔ Los valores *p* muy próximos al valor límite (0,05) se consideran marginales (caben ambas posibilidades). Debes indicar siempre el valor *p* para que quienes lean tus resultados puedan extraer sus propias conclusiones.

Por ejemplo, imagina que una pizzería dice que entrega las pizzas en treinta minutos o menos, en promedio, pero tú crees que tardan más. Realizas un contraste de hipótesis porque crees que la hipótesis nula H_0 (según la cual el tiempo medio de entrega es de treinta minutos como máximo) es incorrecta. Tu hipótesis alternativa (H_a) es que el tiempo medio de entrega es superior a treinta minutos. Tomas una muestra aleatoria de varios tiempos de entrega y sometes los datos al contraste de hipótesis, y el valor *p* resulta ser 0,001, muy por debajo de 0,05. Concluyes que la pizzería está equivocada; el tiempo de entrega de las pizzas supera los treinta minutos en promedio y quieres saber qué piensan hacer al respecto (naturalmente, también podrías haberte equivocado por haber incluido en tu muestra por puro azar un número inusualmente alto de pizzas entregadas con retraso, pero eso no se lo creen ni los de la pizzería). En el capítulo 14 encontrarás más información sobre los valores *p*.

Significación estadística

Cuando se recopilan datos para realizar un contraste de hipótesis, el investigador generalmente busca algo que se salga de lo normal (por desgracia, las investigaciones que se limitan a confirmar algo que ya era conocido no generan titulares). Los estadísticos utilizan los contrastes de hipótesis (ver el capítulo 14) para medir cuánto se sale de lo normal un determinado resultado. Para ello, consideran que un resultado es *estadísticamente significativo* cuando existe una probabilidad muy pequeña de que haya ocurrido por mero azar, y proporcionan un número llamado valor *p* para reflejar dicha probabilidad. (Los valores *p* se tratan en el apartado anterior.)

Por ejemplo, si se comprueba que un fármaco es más eficaz que el tratamiento actual para el cáncer de mama, los investigadores dicen que el nuevo fármaco supone una mejora estadísticamente significativa en la tasa de supervivencia de las pacientes con cáncer de mama. Esto significa que, a partir de los datos obtenidos, la diferencia entre los resultados generales de las pacientes que tomaron el nuevo fármaco y los resultados de las que tomaron el tratamiento anterior es tan grande que sería muy difícil decir que obedece a una simple coincidencia. De todos modos, ándate con cuidado: no puedes decir que esos resultados sean aplicables a todas las personas ni de igual modo a todas las personas. En el capítulo 14 trato más a fondo la cuestión de la significación estadística.

Cuando oigas decir que los resultados de un estudio son estadísticamente significativos, no des por sentado que esos resultados son importantes. *Estadísticamente significativo* significa que los resultados se salen de lo normal, pero eso no siempre significa que sean importantes. Por ejemplo, ¿te emocionaría mucho descubrir que los gatos mueven la cola más a

menudo cuando están tumbados al sol que cuando están a la sombra, y que esos resultados son estadísticamente significativos? ¡Ese resultado no le importa siquiera al gato, y mucho menos a ti!

A veces los estadísticos llegan a la conclusión equivocada sobre la hipótesis nula porque la muestra no representa a la población (aunque eso ocurra por casualidad). Por ejemplo, un efecto positivo experimentado por una muestra de personas que hayan tomado un nuevo tratamiento puede haberse debido a un golpe de suerte; o siguiendo el ejemplo del apartado anterior, es posible que la pizzería sí entregue las pizzas a tiempo y tú, por mala suerte, hayas elegido una muestra de pizzas entregadas con retraso. No obstante, lo bonito de la investigación es que, en cuanto alguien lanza un comunicado de prensa diciendo que ha descubierto algo significativo, todo el mundo intenta reproducir esos resultados, y, si no es posible reproducirlos, probablemente sea porque los resultados originales eran incorrectos por alguna razón (puede incluso que por puro azar). Por desgracia, los comunicados de prensa que anuncian "avances revolucionarios" tienen mucha repercusión en los medios, mientras que los estudios posteriores que refutan esos resultados casi nunca aparecen en portada.

No te apresures a tomar decisiones por haber obtenido un resultado estadísticamente significativo. En ciencia, un estudio aislado, por extraordinario que sea, generalmente no tiene tanto valor como un conjunto de pruebas acumuladas a lo largo de mucho tiempo, junto con varios estudios de seguimiento bien diseñados. Cuando te hablen de algún logro extraordinario, acéptalo con reservas y espera a que salgan a la luz nuevas investigaciones antes de utilizar la información de un único estudio para tomar decisiones importantes que afecten a tu vida. Puede que los resultados no puedan reproducirse o que, incluso en tal caso, no puedas saber si son aplicables a todo el mundo.

Correlación y causalidad

De todos los equívocos que pueden ocurrir en el campo de la estadística, el que quizá resulta más problemático es el mal uso de los conceptos de correlación y causalidad.

La *correlación*, como término estadístico, es la medida en que dos variables numéricas presentan una relación lineal (es decir, una relación que aumenta o disminuye a un ritmo constante). Aquí tienes tres ejemplos de variables correlacionadas:

- ✔ El número de chirridos que emite un grillo en un minuto está estrechamente relacionado con la temperatura: cuando hace frío, el grillo canta menos veces, y a medida que aumenta la temperatura canta con una frecuencia cada vez mayor. En términos estadísticos, decimos que el número de chirridos y la temperatura presentan una fuerte correlación positiva.
- ✔ Se ha encontrado una relación entre el número de delitos (por habitante) y el número de policías en una determinada zona. Cuando hay más policías patrullando en la zona, tiende a haber menos delitos y viceversa, cuando hay menos policías presentes en la zona, generalmente se cometen más delitos. En términos estadísticos decimos que el número de policías y el número de delitos presentan una fuerte correlación negativa.
- ✔ El consumo de helado (litros por persona) y el número de homicidios en Nueva York presentan una correlación positiva. Es decir, a medida que aumentan las ventas de helado por habitante, aumenta también el número de homicidios. ¡Por extraño que parezca, es cierto!

Pero la correlación como estadístico no puede explicar por qué existe una relación entre dos variables *x* e *y*; tan sólo nos dice que existe.

La *causalidad* va un paso más allá que la correlación y significa que un cambio en el valor de la variable *x* causará un cambio en el valor de la variable *y*. Este paso adelante se da más veces de las debidas en las investigaciones, en los medios de comunicación y en el consumo público de resultados estadísticos. Por ejemplo, no puedes decir que el consumo de helado causa un incremento en la tasa de homicidios sólo porque ambas cosas estén correlacionadas. De hecho, el estudio mostraba que la temperatura presenta una correlación positiva con las ventas de helado y también con los homicidios. (En el capítulo 18 profundizo sobre la correlación y la causalidad.) ¿Cuándo puedes decir que existe una relación de causalidad? El caso más claro es cuando se lleva a cabo un experimento bien diseñado que descarta otros factores que podrían estar relacionados con los resultados. (En el capítulo 17 encontrarás información sobre experimentos que revelan una relación causa-efecto.)

Es posible que al observar una correlación sientas el deseo de anunciar una relación causa-efecto; los investigadores, los medios y el público en general lo hacen continuamente. Sin embargo, antes de extraer ninguna conclusión averigua cómo se han recopilado los datos y espera a ver si otros investigadores logran reproducir los resultados (es lo primero que intentan cuando ven que un resultado “revolucionario” de un colega se convierte en noticia de portada).

Parte II

Nociones de cálculo numérico

—HE EVALUADO NUESTRO ÚLTIMO GRÁFICO DE TARTA. YO DIRÍA QUE ES DE MORAS.

En esta parte...

Cálculo numérico: es un trabajo sucio pero alguien tiene que hacerlo. ¿Por qué no tú? Aunque no te vayan mucho los números ni los cálculos, el planteamiento progresivo de esta parte del libro quizá sea lo que necesitas para ganar confianza en la elaboración y comprensión profunda de estadísticos.

En esta parte te explico los fundamentos del cálculo numérico, desde la confección e interpretación de diagramas y gráficos hasta el cómputo y la comprensión de medias, medianas, desviaciones estándares y más. Además, te enseño a hacer un análisis crítico de la información estadística aportada por otras personas y a ver la auténtica verdad que se oculta tras los datos.

Capítulo 5

Medias, medianas y más

En este capítulo

- Resumir datos eficazmente
- Interpretar las estadísticas más utilizadas
- Saber ver lo que dicen y no dicen las estadísticas

Todo conjunto de datos tiene una historia detrás que es posible descubrir y dar a conocer si los estadísticos se utilizan como es debido. Los estadísticos mal utilizados pueden contarte una historia distinta, o sólo una parte de ella, de manera que es muy importante conocer la manera correcta de tomar decisiones sobre la información recibida.

Un *estadístico descriptivo* (o *estadístico*, para abreviar) es un número que resume o describe alguna característica de un conjunto de datos. En este capítulo verás algunos de los estadísticos descriptivos más comunes y la manera en que se utilizan, y aprenderás a calcularlos, interpretarlos y juntarlos todos para obtener una visión clara de un conjunto de datos. Asimismo, descubrirás lo que esos estadísticos dicen y no dicen sobre los datos.

Resumir datos con estadísticos descriptivos

Los estadísticos descriptivos cogen un conjunto de datos y lo reducen a la información básica. Los datos resumidos suelen utilizarse para proporcionar a la gente información que sea fácil de entender y ayude a responder sus preguntas. Imagina que se te acerca tu jefe y te pregunta: "¿Cuál es nuestra base de clientes en la actualidad y quién compra nuestros productos?". ¿Cómo responderías a esa pregunta? ¿Con una larga lista de números incomprensibles que casi con toda seguridad le dejarían medio aturdido? Probablemente no. Quieres unas estadísticas claras y concisas que resuman la base de clientes para que tu jefe vea lo brillante que eres y luego te envíe a recopilar más datos incluso, a ver si descubre la manera de incluir a más personas en esa base de clientes (eso es lo que sacas por ser tan eficiente). Los datos también se resumen con otros fines. Una vez que se han recopilado todos los datos de una encuesta o de otro tipo de estudio, el siguiente paso para el investigador es encontrarles algún sentido. Generalmente lo primero que hacen los investigadores es calcular algunos estadísticos básicos para hacerse una idea aproximada de lo que significan los resultados. Más adelante los investigadores pueden hacer más análisis para formular o comprobar afirmaciones sobre la población de la que proceden los datos, estimar ciertas características sobre dicha población (por ejemplo la media), buscar vínculos entre variables medidas, etc.

Otra parte importante de la investigación consiste en comunicar los resultados, no sólo a otros investigadores, sino también a los medios y al público en general. Por mucho que la comunidad científica pueda estar deseosa de echar mano a los complejos análisis realizados con un conjunto de datos, el público en general no está preparado ni tiene interés alguno en tal cosa. ¿Qué es lo que el público quiere? Información básica. Los estadísticos que comunican una idea de forma clara y concisa son los que normalmente se utilizan para transmitir información a los medios y al público.

Si realmente necesitas extraer más información de los datos, un análisis estadístico rápido no es suficiente. En el campo de la estadística no se cumple la máxima de que menos es más, y a veces con todo el jaleo puedes perder de vista la auténtica verdad que se oculta tras los datos. Para ser un consumidor de estadística bien informado tienes que reflexionar sobre qué estadísticas te están dando, qué significan realmente y qué información te falta. En este capítulo profundizo en estas cuestiones.

Cálculos con datos categóricos: tablas y porcentajes

Los *datos categóricos* (también llamados *datos cualitativos*) expresan cualidades o características de la persona, como por ejemplo su color de ojos, sexo, filiación política u opinión sobre una determinada cuestión (utilizando categorías como "a favor", "en contra" o "sin opinión"). Generalmente se enmarcan fácilmente en grupos o categorías; por ejemplo, en Estados Unidos hay cuatro grandes grupos para la variable "partido político": demócratas, republicanos, independientes y otros. Los datos categóricos generalmente proceden de encuestas, aunque también pueden recopilarse en experimentos. Por ejemplo, en una prueba experimental de un nuevo tratamiento médico los investigadores pueden utilizar tres categorías para valorar los resultados del experimento: ¿el paciente mejoró, empeoró o continuó igual mientras se le administraba el tratamiento? Los datos categóricos a menudo se resumen comunicando el porcentaje de personas que entran en cada categoría. Por ejemplo, los encuestadores pueden comunicar las estadísticas de filiación política dando el porcentaje de republicanos, demócratas, independientes y otros. Para calcular el porcentaje de personas que pertenecen a una determinada categoría, averigua el número de personas de esa categoría, divide por el número total de personas del estudio y luego multiplica por 100. Por ejemplo, si en una encuesta de 2.000 adolescentes había 1.200 chicas y 800 chicos, los porcentajes resultantes serían (1.200/2.000)×100=60% de chicas y (800/2.000)×100 =40% de chicos.

Los datos categóricos aún pueden desglosarse más creando tablas de contingencia. Las *tablas de contingencia* (también llamadas *tablas de doble entrada*) son tablas con filas y columnas que resumen la información de dos variables categóricas a la vez, por ejemplo el sexo y la filiación

política, para que puedas ver (o calcular fácilmente) el porcentaje de personas en cada combinación de categorías y realizar comparaciones entre grupos.

Por ejemplo, si tienes datos sobre el sexo y el partido político de las personas que han respondido a tu encuesta, podrás saber el porcentaje de mujeres republicanas, hombres republicanos, mujeres demócratas, hombres demócratas, etc. En este ejemplo, el número total de combinaciones posibles en la tabla sería 2×4=8, el número total de categorías de sexo multiplicado por el número total de categorías de filiación política. (Las tablas de contingencia se tratan a fondo en el capítulo 19, con algunos añadidos.)

Los gobiernos habitualmente calculan y resumen montañas de datos categóricos utilizando tablas de doble entrada. Por ejemplo, la tabla 5-1 recoge datos típicos de edad y sexo comunicados por la Oficina del Censo de Estados Unidos para una encuesta realizada en 2009 (normalmente la edad se consideraría una variable numérica, pero, debido a la forma en que la comunica el gobierno, dividida en categorías, pasa a ser una variable categórica).

Trabajando con los números de la tabla 5-1 puedes examinar muchas facetas diferentes de la población de Estados Unidos. Si te fijas en el sexo, por ejemplo, verás que las mujeres superan en número a los varones por muy poco —en 2009 las mujeres eran el 50,67% de la población (divide el número total de mujeres por el tamaño total de la población y multiplica por 100) y los varones el 49,33% (divide el número total de varones por el tamaño total de la población y multiplica por 100)—. También puedes fijarte en la edad: el 6,94% de toda la población tenía menos de cinco años (divide el número total de personas con menos de cinco años por el tamaño total de la población y multiplica por 100). El grupo más numeroso era el de personas entre cuarenta y cinco y cuarenta y nueve años, que representaban el 7,44% de la población.

Tabla 5-1. Población de Estados Unidos clasificada por edades y sexos (2009)

Grupo de edad	*Ambos sexos*	*%*	*Varones*	*%*	*Mujeres*	*%*
Menos de 5	21.299.656	6,94	10.887.008	7,19	10.412.648	6,69
5-9	20.609.634	6,71	10.535.900	6,96	10.073.734	6,48
10-14	19.973.564	6,51	10.222.522	6,75	9.751.042	6,27
15-19	21.537.837	7,02	11.051.289	7,30	10.486.548	6,74
20-24	21.539.559	7,02	11.093.552	7,32	10.446.007	6,72
25-29	21.677.719	7,06	11.115.560	7,34	10.562.159	6,79
30-34	19.888.603	6,48	10.107.974	6,67	9.780.629	6,29
35-39	20.538.351	6,69	10.353.016	6,84	10.185.335	6,55
40-44	20.991.605	6,84	10.504.139	6,94	10.487.466	6,74
45-49	22.831.092	7,44	11.295.524	7,46	11.535.568	7,42
50-54	21.761.391	7,09	10.677.847	7,05	11.083.544	7,13
55-59	18.975.026	6,18	9.204.666	6,08	9.770.360	6,28
60-64	15.811.923	5,15	7.576.933	5,00	8.234.990	5,29
65-69	11.784.320	3,84	5.511.164	3,64	6.273.156	4,03
70-74	9.007.747	2,93	4.082.226	2,70	4.925.521	3,17
75-79	7.325.528	2,39	3.149.236	2,08	4.176.292	2,68
80-84	5.822.334	1,90	2.298.260	1,52	3.524.074	2,27
85-89	3.662.397	1,19	1.266.899	0,84	2.395.498	1,54
90-94	1.502.263	0,49	424.882	0,28	1.077.381	0,69
95-99	401.977	0,13	82.135	0,05	319.842	0,21
100+	64.024	0,02	8758	0,01	55.266	0,04
Total	307.006.550	100,00	151.449.490	100,00	155.557.060	100,00

A continuación puedes buscar una posible relación entre sexo y edad comparando varias partes de la tabla. Por ejemplo, puedes comparar el porcentaje de mujeres y varones que hay en el grupo de ochenta años o más. Pero como estos datos se facilitan en incrementos de cinco años, tendrás que hacer algunos cálculos para obtener una respuesta. El porcentaje de la población que son mujeres y tienen ochenta años o más (mirando la columna 7 de la tabla 5-1) es 2,27+1,54+0,69+0,21+0,04=4,75%. El porcentaje de hombres de ochenta años o más (mirando la columna 5 de la tabla 5-1) es 1,52+0,84+0,28+0,05+0,01=2,70%. Por tanto, el grupo de mujeres de ochenta años o más es aproximadamente el 76% más numeroso que el de varones de esa misma edad (porque [4,75–2,70]/2,70=0,76).

Estos datos confirman la creencia ampliamente aceptada de que las mujeres suelen vivir más que los hombres. Sin embargo, la diferencia entre hombres y mujeres es cada vez menor. Según la Oficina del Censo de Estados Unidos, en 2001 el 4,36% de las mujeres tenían ochenta años o más, frente al 2,31% de varones. En 2001 las mujeres pertenecientes a ese grupo de edad superaban a los varones en el 89% (observa que [4,36–2,31] /2,31=0,89).

Una vez que tengas las tablas de doble entrada con el desglose de dos variables categóricas, puedes realizar contrastes de hipótesis para determinar si existe una relación o vínculo significativo entre esas dos variables, teniendo en cuenta que los datos varían de una muestra a otra.

En el capítulo 14 encontrarás información exhaustiva sobre los contrastes de hipótesis.

Medir el centro con la media y la mediana

En el caso de los *datos numéricos*, las características mensurables como la altura, el peso, el coeficiente intelectual, la edad o los ingresos se representan con números que tienen sentido en el contexto del problema (por ejemplo en unidades de centímetros, dólares o personas). Como los datos tienen un significado numérico, se pueden resumir de más maneras que los datos categóricos. El modo más frecuente de resumir un conjunto de datos numéricos es indicar dónde está el centro y una forma de pensar dónde está el centro de un conjunto de datos es preguntarse qué es un valor típico; o bien dónde está el centro de los datos. En realidad el centro de un conjunto de datos puede medirse de distintas formas, y el método elegido puede influir mucho en las conclusiones que se extraigan sobre esos datos. En este apartado hablaré de las formas de medir dónde está el centro.

Calcular la media

Los jugadores de la NBA ganan mucho dinero, ¿verdad? Más de una vez habrás oído decir que jugadores como Kobe Bryant o LeBron James ganan decenas de millones de dólares al año. Pero ¿es eso lo que gana el jugador de la NBA típico? Ya te digo yo que no (aunque tampoco es que me sienta mal por ellos, porque igualmente ganan más pasta de la que tú y yo ganaremos nunca). Decenas de millones de dólares es la cantidad de dinero que puedes exigir cuando eres una superestrella entre las superestrellas, que es precisamente lo que son esos jugadores de élite. Pero volvamos a la pregunta: ¿cuánto gana un jugador de la NBA típico? Una forma de responder es calculando la media (el estadístico más utilizado de todos los tiempos).

La *media* o *promedio* de un conjunto de datos se designa como $\bar{x}$. La fórmula para encontrar la media es la siguiente:

$$\bar{x} = \frac{\sum x_i}{n}$$

en la que cada valor del conjunto de datos se designa mediante una *x* con un subíndice *i* que va desde 1 (el primer número) hasta *n* (el último número).

Así se calcula la media de un conjunto de datos:

1. **Suma todos los números del conjunto de datos.**
2. **Divide por la cantidad de números que haya en el conjunto de datos, *n*.**

La media a la que me refiero aquí es aplicable a una muestra de datos y técnicamente se llama *media muestral*. La media de una población entera se designa con la letra griega μ y se llama *media poblacional*. Para calcularla se suman todos los valores de la población y se divide por el tamaño de la población, designado como *N* (para distinguirlo de un tamaño de muestra, *n*). Generalmente la media poblacional se desconoce y se utiliza una media muestral para estimarla (más/menos un margen de error; más información en el capítulo 13).

Por ejemplo, en la tabla 5-2 aparecen los datos correspondientes al sueldo de los 13 jugadores que ganaron el campeonato de la NBA en 2010 con Los Angeles Lakers.

Tabla 5-2. Sueldos percibidos por los jugadores de Los Angeles Lakers (2009-2010)

Jugador	*Sueldo (en dólares)*
Kobe Bryant	23.034.375
Pau Gasol	16.452.000
Andrew Bynum	12.526.998
Lamar Odom	7.500.000
Ron Artest	5.854.000
Adam Morrison	5.257.229
Derek Fisher	5.048.000
Sasha Vujacic	5.000.000
Luke Walton	4.840.000
Shannon Brown	2.000.000
Jordan Farmar	1.947.240
Didier Ilunga-Mbenga	959.111
Josh Powell	959.111
Total	**91.378.064**

La media de todos los sueldos del equipo es 91.378.064 dólares/13 = 7.029.082 dólares (5.452.319 euros). No está mal como sueldo medio, ¿verdad? Sin embargo, el sueldo de Kobe Bryant está muy por encima de los demás (de hecho, fue el segundo jugador mejor pagado de la liga esa temporada, por detrás de Tracy McGrady). Si quitas a Kobe de la ecuación (literalmente), el sueldo medio de todos los jugadores de los Lakers menos Kobe pasa a ser 68.343.689 dólares/12=5.695.307 dólares (o, en euros, la friolera de 4.344.564), alrededor de 1,3 millones menos que antes.

Esta nueva media continúa siendo una barbaridad de dinero, pero es bastante menos que el sueldo medio de todos los jugadores incluido Kobe (los admiradores de este jugador argumentarían que su sueldo está en consonancia con su importancia para el equipo, y otros dirían que nadie se merece tantísimo dinero; esta cuestión no es más que la punta del iceberg de las discusiones interminables que los forofos de los deportes, entre los que me incluyo, mantienen sobre las estadísticas).

En resumidas cuentas: la media no te cuenta siempre toda la verdad. A veces puede ser un poco engañosa, y éste es uno de esos casos. La razón es que todos los años hay unos pocos jugadores de altísimo nivel (como Kobe) que ganan mucho más dinero que los demás, y sus sueldos hacen que suba el salario medio.

Las cifras de un conjunto de datos que son extremadamente altas o extremadamente bajas en comparación con el resto de las cifras se llaman *valores atípicos*. Debido a la forma de hacer los cálculos, los valores atípicos altos tienden a subir la media (como ocurría con el sueldo de Kobe en el ejemplo anterior). Los valores atípicos bajos, por el contrario, bajan la media.

Calcular la mediana

¿Recuerdas cuando en el colegio os ponían un examen sorpresa y, mientras tú y casi todos los de la clase no llegabais al aprobado, siempre había un par de empollones que sacaban un 10? Seguramente tu profesor no presentaba la curva real de puntuaciones porque no quería reflejar el bajo nivel de la mayoría de los alumnos. Es de suponer que utilizaba la media, y la media en este caso no representa realmente lo que un estadístico consideraría la mejor medida del centro de las puntuaciones obtenidas por los alumnos.

¿Qué otro dato puedes aportar, aparte de la media, para mostrar cuál sería el sueldo de un jugador "típico" de la NBA o qué nota sacaba un estudiante "típico" de tu clase? Otro estadístico utilizado para medir el centro de un conjunto de datos es la mediana. Sin embargo, la mediana es un héroe olvidado en el sentido de que se utiliza muchísimo menos de lo que debería, aunque últimamente la gente está empezando a tenerla más en cuenta.

La *mediana* de un conjunto de datos es el valor que se encuentra exactamente en el centro cuando los datos están ordenados. Se designa de distintas maneras: hay quien utiliza la letra *M* y hay quien prefiere el símbolo . Éstos son los pasos para encontrar la mediana de un conjunto de datos:

1. **Ordena las cifras de la más pequeña a la más grande.**
2. **Si el conjunto de datos contiene un número impar de cifras, elige la que esté exactamente en el centro. Ésa es la mediana.**
3. **Si el conjunto de datos contiene un número par de cifras, coge las dos del centro y calcula la media para obtener la mediana.**

Los sueldos de los jugadores de Los Angeles Lakers durante la temporada 2009-2010 (puedes verlos en la tabla 5-2) están ordenados del más bajo (en el extremo inferior) al más alto (en el extremo superior). Como la lista contiene los nombres y los sueldos de 13 jugadores, el sueldo del centro es el séptimo empezando por abajo: Derek Fisher, que ganó 5.048 millones de dólares esa temporada con los Lakers. Derek está en la mediana.

Este sueldo mediano (5,048 millones de dólares) está muy por debajo de la media de 7,029 millones correspondiente al equipo en la temporada 2009-2010. Fíjate que tan sólo 4 de los 13 jugadores estaban por encima del sueldo medio de 7,029 millones de dólares. Puesto que la media incluye los valores atípicos (como el sueldo de Kobe Bryant), el sueldo mediano es más representativo del centro de los sueldos del equipo. A la mediana no le afectan igual que a la media los sueldos estratosféricos de unos pocos jugadores.

Nota: por cierto, el sueldo más bajo en los Lakers durante la temporada 2009-2010 fue de 959.111 dólares (unos 731.858 euros), un montón de pasta si se compara con los ingresos de una persona normal, pero una miseria para ser un jugador de la NBA.

El gobierno de Estados Unidos casi siempre utiliza la mediana para representar el centro en relación con los datos sobre ingresos, porque, como hemos dicho antes, a la mediana no le afectan los valores atípicos. Por ejemplo, la Oficina del Censo de Estados Unidos comunicó que en 2008 los ingresos medianos por unidad familiar ascendieron a 50.233 dólares, mientras que la media se situaba en 68.424 dólares. ¡Menuda diferencia!

Comparar medias y medianas: histogramas

A veces, el debate en torno a la media y la mediana puede ser la mar de interesante. Pongamos que trabajas para un equipo de la NBA que está renegociando los sueldos. Si representas a los propietarios, te interesa que se sepa lo mucho que están cobrando todos y lo mucho que te estás gastando tú, de manera que tendrás en cuenta a las superestrellas y comunicarás el sueldo medio. Pero si estás de parte de los jugadores, te interesará comunicar la mediana, porque es más representativa de lo que cobran los jugadores que están en el centro. El 50% de los jugadores ganan más que la mediana, y el otro 50%, menos que la mediana. Para salir de dudas, lo mejor es calcular y comparar la media y la mediana. Un buen punto de partida es un gráfico que muestre la forma de los datos.

Uno de los gráficos que puedes elaborar para ver la forma de datos numéricos (cuántos valores están cerca/lejos de la media, dónde está el centro, cuántos valores atípicos podría haber) es un histograma. Un *histograma* es una representación gráfica de datos numéricos organizados, donde se muestran grupos de datos y el número o porcentaje de datos que pertenecen a cada grupo. Proporciona una impresión visual del conjunto de datos. (En el capítulo 7 encontrarás más información sobre histogramas y otras formas de representación gráfica.)

Los conjuntos de datos pueden adoptar muchas formas. Aquí tienes las tres que aparecen con más frecuencia en los cursos de introducción a la estadística:

- ✔ Si la mayor parte de los datos están en el lado izquierdo del histograma pero hay unos pocos valores más grandes en la derecha, se dice que los datos están *sesgados a la derecha* (o que tienen *sesgo positivo*.)

  El histograma A de la figura 5-1 muestra un ejemplo de datos sesgados a la derecha. Los pocos valores más grandes hacen que suba la media, pero apenas afectan a la mediana. Por lo tanto, cuando los datos están sesgados a la derecha la media es más grande que la mediana. Un ejemplo serían los sueldos de la NBA.

- ✔ Si la mayor parte de los datos están en el lado derecho pero hay unos pocos valores más pequeños en el lado izquierdo del histograma, los datos están *sesgados a la izquierda* (tienen *sesgo negativo*.)

  El histograma B de la figura 5-1 muestra un ejemplo de datos sesgados a la izquierda. Los pocos valores más pequeños bajan la media, pero influyen muy poco (o nada) en la mediana. Un ejemplo de datos sesgados a la izquierda es el tiempo que tardan los alumnos en hacer un examen; unos pocos terminan en seguida, otros tardan algo más, y la mayoría no entregan hasta el último momento (¡alguno se quedaría el examen para siempre si pudiera!). Cuando los datos están sesgados a la izquierda la media es más pequeña que la mediana.

- ✔ Si los datos son *simétricos*, tienen más o menos la misma forma en ambos lados del centro. Dicho de otro modo: si doblas el histograma por la mitad, los dos lados son casi iguales.

  El histograma C de la figura 5-1 muestra un ejemplo de datos simétricos. Cuando los datos son simétricos, la media y la mediana son muy parecidas.

Si miras el histograma A de la figura 5-1 (cuya forma está sesgada a la derecha), verás que la cola de la gráfica (donde las barras se hacen más cortas) está en el lado derecho, mientras que en el histograma B (cuya forma está sesgada a la izquierda) la cola está en el lado izquierdo. La dirección de la cola de una distribución sesgada indica la dirección del sesgo. Cuando describas una distribución sesgada, debes mencionar siempre la dirección.

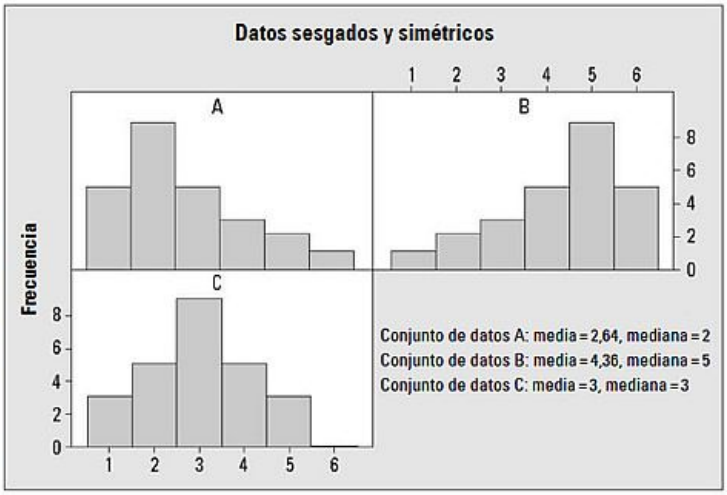


Figura 5-1: A) datos sesgados a la derecha; B) datos sesgados a la izquierda; y C) datos simétricos

El histograma C es simétrico (presenta aproximadamente la misma forma por ambos lados). Sin embargo, no todos los datos simétricos tienen

forma de campana con ocurre en el histograma C. Mientras la forma sea aproximadamente igual en ambos lados, puedes afirmar que es simétrica.

Los valores atípicos afectan a la media de un conjunto de datos, pero no a la mediana. En la jerga técnica, cuando un estadístico no se ve afectado por una determinada característica de los datos (por ejemplo valores atípicos o sesgo), se dice que ese estadístico es *resistente* a esa característica. En este caso la mediana es resistente a los valores atípicos, pero la media no. Si te dan el valor medio de una variable que te interesa, pide también la mediana para poder comparar ambos estadísticos y tener una visión más clara de la situación y de lo que es realmente un valor típico.

Tener en cuenta la variación

Con independencia de las características que estés midiendo, en un conjunto de datos siempre existe variación, porque no todo el mundo va a tener exactamente el mismo valor para cada variable. La variación es lo que convierte a la estadística en lo que es. Por ejemplo, el precio de la vivienda varía de una vivienda a otra, de un año a otro y de un estado a otro. El tiempo que tardas en llegar al trabajo varía de un día a otro. El truco está en saber medir la variación de la mejor manera posible.

Comunicar la desviación estándar

Con diferencia, la medida más común para la variación de datos numéricos es la *desviación estándar*, que mide el grado de concentración de los datos en torno a la media; cuanto más concentrados están, más pequeña es la desviación estándar. Este estadístico no se proporciona tan a menudo como sería deseable, pero cuando sí lo indican, generalmente aparece entre paréntesis: (*s*=2,68).

Calcular la desviación estándar

La fórmula para calcular la desviación estándar de un conjunto de datos *(s)* es

$$s=\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

Para calcular *s*, sigue los pasos siguientes:

1. **Encuentra la media del conjunto de datos, $\bar{x}$.**
2. **Coge cada número del conjunto de datos *(x)* y réstale la media para obtener $(x-\bar{x})$.**
3. **Calcula el cuadrado de cada diferencia, $(x-\bar{x})^2$.**
4. **Suma todos los resultados del paso 3 para obtener la suma de los cuadrados: $\sum(x-\bar{x})^2$.**
5. **Divide la suma de los cuadrados (calculada en el paso 4) por la cantidad de números del conjunto de datos menos uno; es decir, (*n*–1). Ahora tienes:**

$$\frac{\sum(x-\bar{x})^2}{n-1}$$

6. **Calcula la raíz cuadrada para obtener**

$$s=\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

que es la desviación estándar de la muestra, *s*. ¡Buf!

Al final del paso 5 obtienes un estadístico llamado *variancia muestral*, designada como s^2. La variancia es otra forma de medir la variación en un conjunto de datos; la pega es que se indica en unidades cuadradas. Si los datos están en dólares, por ejemplo, la variancia estaría en dólares cuadrados (lo cual no tiene sentido). Por eso se añade el paso 6. La desviación estándar tiene las mismas unidades que los datos originales.

Veamos un ejemplo. Imagina que tienes cuatro notas de examen: 1, 3, 5 y 7. La media es 16/4=4 puntos. Si restas la media a cada número, obtienes (1–4)=–3, (3–4)=– 1, (5–4)=+1, y (7–4)=+3. Al calcular el cuadrado de cada resultado obtienes 9, 1, 1 y 9. Si los sumas todos, el total es 20. En este ejemplo, *n*=4 y, por tanto, *n*–1=3, así que divides 20 entre 3 y te da 6,67. Las unidades en este caso son "puntos al cuadrado", lo cual, obviamente, no tiene sentido. Por último, calculas la raíz cuadrada de 6,67, que es 2,58. La desviación estándar para estas cuatro notas de examen es 2,58 puntos.

Como para calcular la desviación estándar hay que seguir muchos pasos, generalmente se utiliza un ordenador. No obstante, conocer la forma de calcular la desviación estándar te ayuda a interpretar mejor este estadístico y puede ayudarte a saber cuándo podrían estar dándote un valor incorrecto.

En la fórmula para calcular *s*, los estadísticos dividen por *n*–1 (en lugar de por *n*) para que los resultados tengan unas propiedades que operan en un plano teórico que queda fuera del alcance de este libro (no es que sea *la dimensión desconocida* pero casi; créeme, por ahora no te hace falta saber más).

La desviación estándar de una población entera de datos se designa con la letra griega σ. Cuando utilizo el término *desviación estándar* me refiero a *s*, la desviación estándar de la muestra (si en algún momento me refiero a la desviación estándar de la población, ya lo diré expresamente).

Interpretar la desviación estándar

La desviación estándar puede ser difícil de interpretar como número aislado. Básicamente, una desviación estándar pequeña significa que la mayoría de los valores del conjunto de datos están próximos a la media de ese conjunto, y una desviación estándar grande significa que la mayoría de los valores del conjunto de datos están más alejados de la media.

Una desviación estándar pequeña puede ser deseable en determinadas situaciones en que los resultados están restringidos, por ejemplo en la cadena de fabricación de productos y el control de calidad. Si una pieza concreta de un vehículo debe tener 2 centímetros de diámetro para que encaje correctamente, más vale que la desviación estándar durante el proceso de fabricación no sea muy grande. En este caso una desviación estándar grande se traduciría en un montón de piezas tiradas a la basura por no encajar bien (la alternativa sería dejar que los coches circularan con problemas).

Pero en situaciones en que te limitas a observar y registrar datos, una desviación estándar grande no es necesariamente algo malo; simplemente refleja un alto grado de variación en el grupo que se está estudiando. Por ejemplo, en el conjunto de los sueldos de todas las personas que trabajan para una determinada empresa, desde el último becario hasta el director general, la desviación estándar puede ser muy grande. Sin embargo, si únicamente te fijas en los becarios, la desviación estándar será mucho menor, porque las personas pertenecientes a ese grupo tienen sueldos menos variables. El segundo conjunto de datos no es mejor, sólo es menos variable.

Igual que ocurre con la media, los valores atípicos afectan a la desviación estándar (después de todo, la fórmula para la desviación estándar incluye la media). Siguiendo con el ejemplo de la NBA, los sueldos de los jugadores de Los Angeles Lakers en la temporada 2009-2010 (mostrados en la tabla 5-2) oscilan entre 23.034.375 dólares (Kobe Bryant) y 959.111 dólares (Didier Ilunga-Mbenga y Josh Powell). ¡Mucha variación, desde luego! La desviación estándar de los sueldos de este equipo es de 6.567.405 dólares; casi tanto como la media. No obstante, como ya habrás adivinado, si quitamos del conjunto de datos el sueldo de Kobe Bryant, la desviación estándar disminuye porque el resto de los sueldos están más concentrados en torno a la media. Entonces la desviación estándar pasa a ser 4.671.508 dólares.

Para saber si una desviación estándar es grande, fíjate en las unidades. Por ejemplo, una desviación estándar de 2 en unidades de años es equivalente a una desviación estándar de 24 en unidades de meses. Asimismo, para tener una buena perspectiva de la variación estándar debes fijarte en el valor de la media. Si el número medio de grupos de noticias en los que un usuario escribe es 5,2 y la desviación estándar es 3,4, la variación es muy grande, en términos relativos. Pero si estás hablando de la edad de los usuarios de grupos de noticias, en cuyo caso la media es de 25,6 años, esa misma desviación estándar de 3,4 sería comparativamente más pequeña.

Comprender las propiedades de la desviación estándar

Aquí tienes algunas propiedades que pueden ayudarte a interpretar una desviación estándar:

- ✔ La desviación estándar nunca puede ser un número negativo, debido a la forma en que se calcula y al hecho de que mide una distancia (las distancias nunca son números negativos).
- ✔ El valor más pequeño posible para la desviación estándar es 0, y eso sólo ocurre en situaciones artificiales donde todos los números del conjunto de datos son iguales (no hay desviación).
- ✔ La desviación estándar se ve afectada por los valores atípicos (valores extremadamente altos o extremadamente bajos incluidos en el conjunto de datos). El motivo es que la desviación estándar se basa en la distancia a la media (recuerda que la media también se ve afectada por los valores atípicos).
- ✔ La desviación estándar tiene las mismas unidades que los datos originales.

En defensa de la desviación estándar

La desviación estándar es un estadístico de uso habitual, pero a menudo no recibe toda la atención que se merece. Aunque la media y la mediana están muy presentes en los medios de comunicación, es raro verlas acompañadas de algún valor que mida la diversidad del conjunto de datos, de manera que sólo te ofrecen una parte de la información. De hecho, podrías estar perdiéndote la parte más interesante.

Si no te dan la desviación estándar, no puedes saber si los datos están próximos a la media (como ocurre con los diámetros de las piezas de vehículo que salen de una cinta transportadora cuando todo funciona correctamente) o si están repartidos en un intervalo más amplio (como los precios de la vivienda y los ingresos por familia en Estados Unidos).

Por ejemplo, si alguien te dijera que el sueldo inicial medio en la empresa Statistix es de 70.000 dólares, posiblemente pensarías que eso es un montón de pasta. Pero que la desviación estándar para los sueldos iniciales en Statistix sea de 20.000 dólares significa que existe mucha variación en lo que respecta al dinero que puedes ganar, de manera que el sueldo inicial medio de 70.000 dólares ya no es tan informativo como parecía, ¿no?

Por otro lado, si la desviación estándar fuera de sólo 5.000 dólares, tendrías una idea mucho más clara de lo que podrías cobrar al empezar a trabajar en esa empresa. ¿Qué opción resulta más atractiva? Eso es algo que debe decidir cada cual, pero no cabe duda de que será una decisión mucho más informada una vez seas consciente de que la desviación estándar es un factor importante.

Sin la desviación estándar no puedes comparar bien dos conjuntos de datos. Imagina que dos conjuntos de datos tienen la misma media. ¿Significa eso que son exactamente iguales? En absoluto. Por ejemplo, los conjuntos de datos 199, 200, 201, por un lado, y 0, 200, 400, por otro, tienen la misma media (200) pero presentan desviaciones estándares muy diferentes. El primer conjunto de datos tiene una desviación estándar muy pequeña (s=1) en comparación con la del otro conjunto (s=200).

Las referencias a la desviación estándar quizá acaben generalizándose en los medios de comunicación cuando la gente (como tú, por ejemplo) descubra lo que este estadístico puede decir sobre un conjunto de resultados y entonces empiece a pedirla. Por otro lado, también es probable que la desviación estándar se comunique y se utilice cada vez más en el terreno profesional.

Fuera de rango

El rango es otro estadístico utilizado para medir la diversidad en un conjunto de datos. El *rango* es la diferencia entre el valor más grande y el valor más pequeño del conjunto de datos. Es muy fácil de calcular; sólo tienes que poner los números en orden (del más pequeño al más grande) y hacer una resta. A lo mejor por eso el rango se utiliza tanto; desde luego no es por su valor interpretativo.

El rango de un conjunto de datos apenas transmite alguna información. Depende de sólo dos números del conjunto de datos, y ambos pueden reflejar valores extremos (valores atípicos). Yo te aconsejo que te olvides del rango y calcules la desviación estándar, ya que, al tener en cuenta todos los valores, es una medida mucho más informativa de la variación existente en el conjunto de datos. O también puedes calcular otro estadístico llamado *rango intercuartílico*, que es similar al rango pero con una diferencia importante: elimina los problemas causados por los valores atípicos y el sesgo al fijarse únicamente en el 50% central de los datos y

calcular el rango de esos valores. Encontrarás más información al respecto en el apartado "Consideraciones sobre el rango intercuartílico", al final de este capítulo.

La regla empírica (68-95-99,7)

Proporcionar una medida de tendencia central (por ejemplo la media o la mediana) junto con una medida de variación (por ejemplo la desviación estándar o el rango intercuartílico) es una buena forma de describir los valores en una población. En el caso de que la curva de los datos tenga forma de campana (es decir, los datos tengan una distribución normal, de la que te hablo en el capítulo 9), la mejor combinación es la media poblacional y la desviación estándar, y existe una regla especial que relaciona ambos estadísticos para obtener información bastante detallada sobre el total de la población.

La *regla empírica* dice que si una población tiene una distribución normal, con una media poblacional μ y una desviación estándar σ, entonces:

- ✔ Aproximadamente el 68% de los valores se encuentran a una desviación estándar de la media (es decir, entre la media menos una vez la desviación estándar, y la media más una vez la desviación estándar). En notación estadística, esto se representa como $\mu \pm 1\sigma$.
- ✔ Aproximadamente el 95% de los valores se encuentran a dos desviaciones estándares de la media (es decir, entre la media menos dos veces la desviación estándar, y la media más dos veces la desviación estándar). La notación estadística para esto es $\mu \pm 2\sigma$.
- ✔ Aproximadamente el 99,7% de los valores se encuentran a tres desviaciones estándares de la media (es decir, entre la media menos tres veces la desviación estándar, y la media más tres veces la desviación estándar). Los estadísticos utilizan la siguiente notación para representar esto: $\mu \pm 3\sigma$.

La regla empírica también se conoce como *regla 68-95-99,7*, en alusión a estas tres propiedades. Sirve para describir una población, en lugar de una muestra, pero también puedes utilizarla como ayuda para decidir si una muestra de datos procede de una distribución normal. Si la muestra es suficientemente grande y ves que su histograma tiene forma acampanada, puedes comprobar si los datos cumplen los porcentajes 68-95-99,7. En caso afirmativo, es razonable inferir que los datos proceden de una distribución normal. Esto es importante porque la distribución normal tiene muchas ventajas, como podrás ver en el capítulo 9.

La figura 5-2 muestra los tres componentes de la regla empírica. La razón de que haya tantos valores (el 68% aproximadamente) a menos de una desviación estándar de la media es que, cuando los datos tienen forma de campana, la mayoría de los valores se amontonan en el centro, cerca de la media (como puedes ver en la figura 5-2).

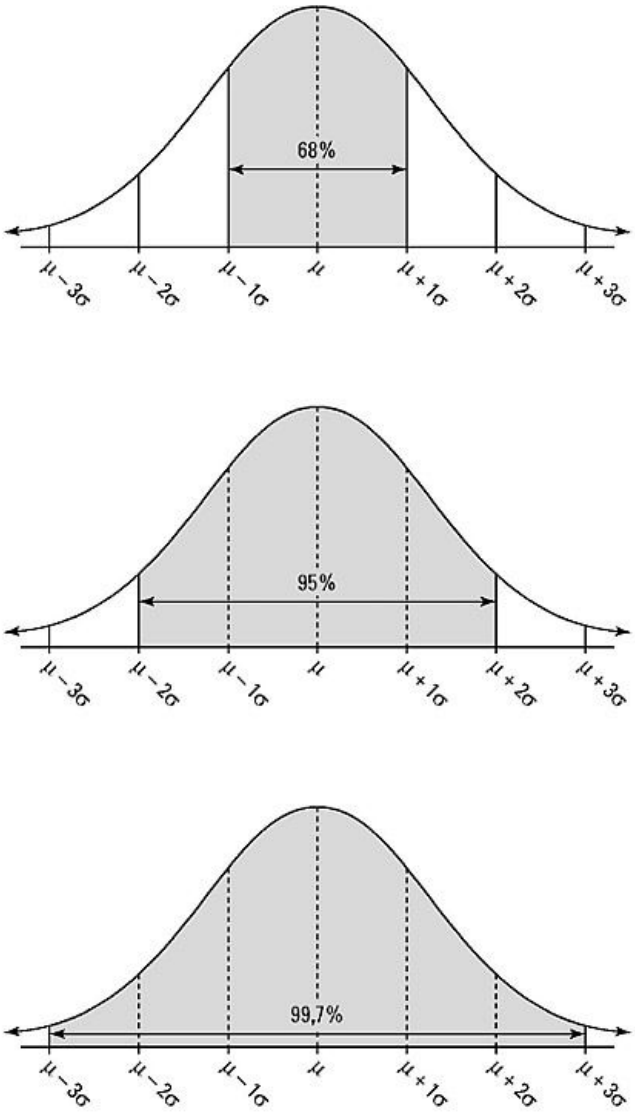


Figura 5-2: La regla empírica (68 %, 95 % y 99,7 %)

Al añadir otra desviación estándar en ambos lados de la media, el porcentaje aumenta de 68 a 95, lo cual supone un salto importante y te da una idea clara de dónde se encuentran la "mayoría" de los datos. La mayoría de los investigadores prefieren un rango del 95% (en lugar del 99,7%) para comunicar sus resultados, ya que no merece la pena aumentar el rango a tres desviaciones estándares por cada lado de la media (en lugar de dos) sólo para incluir ese último 4,7% de los valores.

La regla empírica te dice aproximadamente qué porcentaje de valores se encuentran dentro de un determinado rango respecto de la media, y hago especial hincapié en la palabra "aproximadamente". Estos resultados no son más que aproximaciones, y sólo son aplicables si los datos siguen una distribución normal. No obstante, la regla empírica es un resultado importante en estadística porque la idea de usar dos desviaciones estándares para abarcar el 95% de los valores aparece a menudo en relación con intervalos de confianza y contrastes de hipótesis. (Los intervalos de confianza y los contrastes de hipótesis se tratan en los capítulos 13 y 14, respectivamente.)

Aquí tienes un ejemplo de cómo utilizar la regla empírica para describir mejor una población cuyos valores presentan una distribución normal: en un estudio de cómo la gente hace amistades en el ciberespacio utilizando los grupos de noticias, se decía que los usuarios de un determinado grupo de noticias tenían una edad media de 31,65 años, con una desviación estándar de 8,61 años. Supongamos que los datos se representaron gráficamente con un histograma y seguían una curva en gaussiana similar a la que aparece en la figura 5-2.

Según la regla empírica, aproximadamente el 68% de los usuarios de grupos de noticias tenían edades que se encontraban a una desviación estándar (8,61 años) de la media (31,65 años). Por lo tanto, aproximadamente el 68% de los usuarios tenían entre 31,65–8,61 años y 31,65+8,61 años, es decir, entre 23,04 y 40,26 años. Aproximadamente el 95% de los usuarios de grupos de noticias tenían entre 31,65–2(8,61) años y 31,65+2(8,61) años, es decir, entre 14,43 y 48,87 años. Por último, aproximadamente el 99,7% de los usuarios de grupos de noticias tenían entre 31,65–3(8,61) años y 31,65+3(8,61) años, es decir, entre 5,82 y 57,48 años.

Esta aplicación de la regla proporciona una visión mucho más clara de lo que está ocurriendo en este conjunto de datos que si te limitas a considerar la media, ¿no crees? Como puedes ver, la media y la desviación estándar juntas añaden valor a los resultados. Pasar estos valores por el tamiz de la regla empírica te permite conocer los rangos de la "mayoría" de los datos.

Recuerda: la condición para poder utilizar la regla empírica es que los datos tengan una distribución normal. Si no es así (o si no sabes cuál es la forma de la curva), no puedes utilizar esta regla. Para describir los datos en esos casos puedes utilizar percentiles, que representan ciertos puntos de corte en los datos. (Ver el último apartado: "Hacer un resumen pentanumérico".)

Medir la posición relativa con percentiles

Algunas veces los valores exactos de la media, la mediana y la desviación estándar carecen de importancia, y lo único que interesa es saber dónde te encuentras con respecto al resto de la manada. En esos casos necesitas un estadístico que te diga la posición relativa, y ese estadístico se llama percentil. El *percentil k* es un número del conjunto de datos que divide los datos en dos partes: la parte inferior contiene el *k*% de los datos, y la parte superior contiene el resto de los datos (es decir, el [100–*k*]%, ya que la totalidad de los datos es el 100%). ***Nota:*** *k* es cualquier número entre 1 y 100.

La mediana es el percentil 50: el punto en que el 50% de los datos están por debajo y el otro 50% están por encima.

En este apartado te enseño a calcular, interpretar y juntar percentiles para descubrir la verdad que se oculta tras un conjunto de datos.

Calcular percentiles

Para calcular el percentil *k* (donde *k* es un número cualquiera entre 1 y 100), sigue los pasos siguientes:

1. **Ordena todos los números del conjunto de datos del más pequeño al más grande.**

2. **Multiplica el porcentaje *k* por la cantidad total de números, *n*.**

3a. **Si el resultado del paso 2 es un número entero, sigue con el paso 4. Si el resultado del paso 2 no es un número entero, redondéalo al siguiente número entero y ve al paso 3b.**

3b. **Cuenta los números del conjunto de datos de izquierda a derecha (del más pequeño al más grande) hasta llegar al valor indicado por el paso 3a. El valor correspondiente del conjunto de datos es el percentil *k*.**

4. **Cuenta los números del conjunto de datos de izquierda a derecha hasta llegar al indicado por el paso 2. El percentil *k* es la media entre ese valor correspondiente del conjunto de datos y el valor inmediatamente superior.**

Por ejemplo, supongamos que tienes 25 notas de examen ordenadas de la más baja a la más alta, de la manera siguiente: 43, 54, 56, 61, 62, 66, 68, 69, 69, 70, 71, 72, 77, 78, 79, 85, 87, 88, 89, 93, 95, 96, 98, 99, 99. Para encontrar el percentil 90 de estas notas (ordenadas), empieza multiplicando 90por el número total de notas, es decir: 90×25=0,90×25=22,5. Si redondeas al siguiente número entero, te da 23.

Cuenta de izquierda a derecha (del número más pequeño al número más grande del conjunto de datos) hasta llegar al 23.º número del conjunto de datos. Ese número es el 98, y es el percentil 90 de este conjunto de datos.

Ahora pongamos que quieres encontrar el percentil 20. Empieza calculando 0,20×25=5; este número es entero, de manera que saltas del paso 3a al paso 4, que nos dice que el percentil 20 es la media entre el 5.° y 6.° números del conjunto de datos ordenados (62 y 66). Así pues, el percentil 20 es (62+66)/2=64. La mediana (el percentil 50) de las notas de examen es la 13.ª nota: 77.

No existe una fórmula única y definitiva para calcular percentiles. La fórmula que indico está diseñada para facilitar y hacer más intuitivo el cálculo del percentil, sobre todo si haces las cuentas a mano; no obstante, para trabajar con un ordenador se utilizan otras fórmulas. Los resultados obtenidos con otros métodos pueden diferir, pero no mucho.

Interpretar percentiles

Los percentiles indican la posición relativa de un valor en concreto dentro de un conjunto de datos. Si eso es lo que más te interesa, entonces no importa cuáles sean la media y la desviación estándar del conjunto de datos, y tampoco el valor del dato. Lo que importa es la posición, no en relación con la media, sino en relación con el resto de los valores: y eso es lo que te dice el percentil.

Por ejemplo, en el caso de las notas de examen, ¿qué más da cuál sea la media mientras hayas sacado mejor nota que la mayoría de tus compañeros de clase? Quién sabe, a lo mejor el examen era dificilísimo y una nota de 40 sobre 100 era un magnífico resultado (eso me pasó a mí una vez en clase de matemáticas avanzadas, ¡que no te pase nunca a ti!). En este caso la nota en sí no te dice nada, pero el percentil lo dice todo.

Pongamos que tu nota es mejor que el 90% de las notas de la clase. Eso significa que tu nota está en el percentil 90 (por tanto, k=90), con lo que a lo mejor te ponen un sobresaliente. Por el contrario, si tu nota está en el percentil 10 (algo que a ti jamás te ocurriría porque eres un magnífico estudiante), entonces k=10; eso significa que sólo un 10% del resto de las notas son más bajas que la tuya, y el 90% son más altas; en ese caso está claro que no te van a poner un sobresaliente. Una cosa buena de los percentiles es que tienen una interpretación universal: estar en el percentil 95 significa lo mismo con independencia de si estás hablando de notas de examen o de pesos de paquetes enviados por correo; el percentil 95 significa siempre que el 95% del resto de los valores están por debajo del tuyo, y el 5% están por encima. Esto, además, te permite comparar de manera imparcial dos conjuntos de datos que tengan medias y desviaciones estándares diferentes. Este estadístico iguala el terreno de juego y te ofrece una manera de comparar manzanas y naranjas, por así decirlo.

Un percentil no es un porcentaje; un percentil es un número (o la media entre dos números) del conjunto de datos que indica un punto concreto dentro del recorrido completo de los datos. Pongamos que haces un examen de acceso a la universidad y te dicen que tu nota está en el percentil 80. Eso no significa que hayas respondido correctamente al 80% de las preguntas. Significa que el 80% de las notas de otros estudiantes que se hayan examinado contigo son más bajas que la tuya, y el 20% son más altas que la tuya.

Un percentil alto no siempre es algo bueno. Por ejemplo, si tu ciudad está en el percentil 90 en cuanto a tasa de delincuencia en comparación con otras ciudades del mismo tamaño, significa que el 90% de ciudades similares a la tuya tienen una tasa de delincuencia más baja, lo cual no es bueno para ti. Otro ejemplo son las tarjetas de puntuación de los golfistas; en el golf una puntuación baja es algo bueno, de manera que un resultado que te sitúe en el percentil 80 no te permitirá clasificarte para el circuito de la PGA, y lo digo sin acritud.

Comparar ingresos familiares

El gobierno de Estados Unidos suele incluir percentiles en sus resúmenes de datos. Por ejemplo, según la Oficina del Censo de Estados Unidos, los ingresos medianos (percentil 50) por unidad familiar ascendieron a 42.228 dólares en 2001, y a 50.233 dólares en 2007. La Oficina también proporciona varios percentiles para los ingresos por unidad familiar en cada año, incluidos los percentiles 10, 20, 50, 80, 90 y 95. La tabla 5-3 muestra los valores de cada uno de estos percentiles para los años 2001 y 2007.

Tabla 5-3. Ingresos por unidad familiar en Estados Unidos (2001 frente a 2007)

Percentil	*Ingresos por unidad familiar en 2001 (dólares)*	*Ingresos por unidad familiar en 2007 (dólares)*
10	10.913	12.162
20	17.970	20.291
50	42.228	50.233
80	83.500	100.000
90	116.105	136.000
95	150.499	177.000

Si observas los percentiles de 2001 en la tabla 5-3, verás que las cifras correspondientes a los percentiles más bajos están más próximas entre sí que las correspondientes a los percentiles más altos. La diferencia entre el percentil 20 y el percentil 50 es de 24.000 dólares aproximadamente, mientras que la dispersión entre el percentil 50 y el percentil 80 se acerca a 41.000 dólares. La diferencia entre los percentiles 10 y 50 es de tan sólo 31.000 dólares aproximadamente, mientras que la diferencia entre los percentiles 50 y 90 alcanza los 74.000 dólares.

Los percentiles de 2007 son todos más altos que los percentiles de 2001 (¡lo cual es una buena noticia!), y también presentan una mayor dispersión. Para 2007, hay una diferencia de 30.000 dólares aproximadamente entre los percentiles 20 y 50, y de 50.000 dólares aproximadamente entre los percentiles 50 y 80; y ambas diferencias son mayores que en 2001. De forma similar, el percentil 10 está más alejado del percentil 50 (la diferencia ronda los 38.000 dólares) en 2007 que en 2001, y el percentil 50 está más alejado del 90 (86.000 dólares de diferencia) en 2007 que en 2001. Estos resultados nos dicen que los ingresos han aumentado en general en todos los niveles entre 2001 y 2007, pero la diferencia entre esos niveles es cada vez mayor. Por ejemplo, el percentil 10 de ingresos fue 10.913 dólares en 2001 (puedes verlo en la tabla 5-3) y 12.162 dólares en 2007; esto representa un incremento del 11% aproximadamente (resta uno de otro y divide por 10.913). Ahora

compara los percentiles 95 de 2007 y 2001: el incremento es de casi un 18%. Técnicamente habría que ajustar los valores de 2001 para tener en cuenta la inflación, pero ya ves por dónde van los tiros.

Los cambios en los porcentajes afectan a la variabilidad de un conjunto de datos. Por ejemplo, si los aumentos de sueldo se dan en forma de porcentaje, la diversidad de los sueldos también aumenta; viene a ser la idea de que "los ricos se hacen más ricos". Si a un trabajador que gana 30.000 dólares le conceden un aumento del 10%, pasa a cobrar 33.000 dólares (un aumento de 3.000 dólares); pero si está ganando 300.000 dólares, un aumento del 10% hace que pase a cobrar 330.000 dólares (una diferencia de 30.000 dólares). Por tanto, cuando entres a trabajar en un sitio nuevo negocia el sueldo más alto que puedas, porque los aumentos que te den en el futuro también supondrán una cantidad mayor.

Exámenes y percentiles

Cada año millones de alumnos de secundaria estadounidenses hacen un examen de ámbito nacional que forma parte del proceso de acceso a la universidad (es el equivalente a la selectividad en España). Esta prueba, llamada ACT, está pensada para valorar la capacitación de los futuros estudiantes universitarios en las áreas de inglés, matemáticas, lectura y ciencia. Cada sección de la prueba tiene una calificación máxima de 36 puntos.

Cuando salen las notas de la ACT, no se da la media ni la desviación estándar de las puntuaciones obtenidas por los alumnos (si se hiciera, sería un auténtico lío, porque estos estadísticos pueden cambiar para cada examen, y entonces la gente se quejaría de que un examen es más difícil que otro, cuando en realidad las puntuaciones per se no son relevantes). Para evitar estos problemas, y por otros motivos, los resultados de los exámenes se dan por medio de percentiles.

Generalmente los percentiles se proporcionan en forma de una lista predeterminada. Por ejemplo, la Oficina del Censo de Estados Unidos comunica los percentiles, 10, 20, 50, 80, 90 y 95 para los ingresos por unidad familiar (como puedes ver en la tabla 5-3). La prueba de ACT, sin embargo, utiliza los percentiles de manera diferente. En lugar de comunicar las notas de examen correspondientes a una lista de percentiles preconfeccionada, elaboran una lista con todas las calificaciones posibles y comunican el percentil correspondiente, sea el que sea. De este modo cada cual sabe dónde está; sabiendo tu nota, sólo tienes que encontrar tu percentil.

La tabla 5-4 muestra los percentiles de las puntuaciones obtenidas en las pruebas ACT de matemáticas y lectura en 2009. Para interpretar una puntuación, busca la fila correspondiente a la nota y la columna del área en cuestión (por ejemplo, lectura). En la intersección de la fila y la columna encontrarás el percentil representado por cada puntuación; dicho de otro modo, verás qué porcentaje de los estudiantes que se examinaron sacaron una puntuación más baja.

Tabla 5-4. Percentiles para todas las puntuaciones posibles en las pruebas ACT de matemáticas y lectura

Puntuación ACT	*Percentil en matemáticas*	*Percentil en lectura*
34-36	99	99
33	98	97
32	97	95
31	96	93
30	95	91
29	93	88
28	91	85
27	88	81
26	84	78
25	79	74
24	74	70
23	68	65
22	62	59
21	57	54
20	52	47
19	47	41
18	40	34
17	33	30
16	24	24
15	14	19
14	06	14
13	02	09
12	01	06
11	01	03
1-10	01	01

Por ejemplo, imagina que Andrew ha sacado un 30 en el examen de matemáticas; en la tabla 5-4 busca la fila 30 en la columna de matemáticas y comprueba que su puntuación está en el percentil 95. En otras palabras: el 95% de los estudiantes obtuvieron una puntuación más baja que la suya, y sólo el 5% sacaron más nota que él.

Ahora imagina que Andrew también ha sacado un 30 en el examen de lectura. Que una puntuación de 30 represente el percentil 95 en matemáticas no significa necesariamente que una puntuación de 30 también represente el percentil 95 en lectura (probablemente sea razonable esperar que un número menor de estudiantes consiga un 30 o más en la prueba de matemáticas que en la prueba de lectura).

Para comprobar mi teoría, echa un vistazo a la columna 3 de la tabla 5-4, en la fila correspondiente a una puntuación de 30. Como puedes ver, una puntuación de 30 en el examen de lectura sitúa a Andrew en el percentil 91 (un poco menos que en el examen de matemáticas, pero una buena nota igualmente).

Hacer un resumen pentanumérico

En lugar de comunicar una única medida de tendencia central o una única medida de dispersión, puedes crear un grupo de estadísticos y juntarlos todos para obtener una descripción más detallada de un conjunto de datos. La regla empírica (descrita anteriormente en este mismo capítulo) utiliza la media y la desviación estándar juntas para describir un conjunto de datos con forma de campana. Si los datos no tienen forma de campana, se utiliza otro grupo de estadísticos (basados en percentiles) para proporcionar una visión general sobre ellos. Este método consiste en dividir los datos en cuatro partes (con igual cantidad de datos en cada parte) y comunicar los cinco puntos de corte que separan dichas partes. Estos puntos de corte se representan con cinco estadísticos que describen la manera en que están distribuidos los datos.

El *resumen pentanumérico* es un grupo de cinco estadísticos descriptivos que dividen el conjunto de datos en cuatro partes iguales. Los cinco números de un resumen pentanumérico son:

1. El número *mínimo* (más pequeño) del conjunto de datos.
2. El *percentil 25* (también llamado *primer cuartil*, o *Q*1).
3. La *mediana* (el percentil 50).
4. El *percentil 75* (también llamado *tercer cuartil*, o *Q*3).
5. El número *máximo* (más grande) del conjunto de datos.

Por ejemplo, pongamos que quieres saber el resumen pentanumérico de las 25 notas de examen siguientes (ordenadas): 43, 54, 56, 61, 62, 66, 68, 69, 69, 70, 71, 72, 77, 78, 79, 85, 87, 88, 89, 93, 95, 96, 98, 99, 99. El mínimo es 43, el máximo es 99 y la mediana es el número que está justamente en el centro, 77.

Para encontrar *Q*1 y *Q*3 tienes que seguir los pasos indicados en el apartado "Calcular percentiles", con $n=25$. El paso 1 está hecho porque los datos están ordenados. Para el paso 2, como *Q*1 es el percentil 25, multiplicamos $0{,}25\times25=6{,}25$. Este número no es entero, de manera que, según dice el paso 3a, lo redondeamos al alza hasta 7 y continuamos con el paso 3b.

Según el paso 3b, debes contar de izquierda a derecha en el conjunto de datos hasta llegar al 7.° número, 68; ya tienes *Q*1. Para *Q*3 (el percentil 75) multiplicas $0{,}75\times25=18{,}75$ y redondeas a 19. El 19.º número de la lista es 89, ahí está *Q*3. Si lo juntas todo, te sale que el resumen pentanumérico para estas 25 notas de examen es 43, 68, 77, 89 y 99. La mejor manera de interpretar un resumen pentanumérico es con un diagrama de cajas. Encontrarás más información al respecto en el capítulo 7.

Consideraciones sobre el rango intercuartílico

El propósito del resumen pentanumérico es proporcionar estadísticos descriptivos para el centro, la variación y la posición relativa, todo de una vez. La medida del centro en el resumen pentanumérico es la mediana, mientras que el primer cuartil, la mediana y el tercer cuartil son medidas de la posición relativa.

Para obtener una medida de la variación a partir del resumen pentanumérico, puedes calcular el *rango intercuartílico* (o RIC). El RIC es igual a $Q3-Q1$ (es decir, el percentil 75 menos el percentil 25) y refleja el espacio ocupado por el 50% central de los datos. Si el RIC es pequeño, sabes que muchos datos están próximos a la mediana. Si el RIC es grande, sabes que los datos están más dispersos en torno a la mediana. El RIC del conjunto de datos formado por las notas de examen es $89-68=21$, un número bastante grande teniendo en cuenta que las notas sólo van de 0 a 100.

El rango intercuartílico mide la variación mucho mejor que el rango normal (valor máximo menos valor mínimo; lee el apartado "Fuera de rango" en este mismo capítulo). La razón es que el rango intercuartílico no tiene en cuenta los valores atípicos, sino que los descarta para centrarse exclusivamente en la distancia que ocupan el 50% central de los datos (es decir, el espacio entre los percentiles 25 y 75).

Si se eligen bien y se usan correctamente, los estadísticos descriptivos pueden darte mucha información sobre un conjunto de datos, por ejemplo dónde está el centro, cuál es el grado de diversidad y dónde se encuentra una parte importante de los datos. Sin embargo, los estadísticos descriptivos tampoco lo dicen todo, y en ocasiones pueden resultar equívocos. Estate atento a situaciones en que convendría más utilizar otro estadístico (por ejemplo, la mediana describe el centro de manera más ecuánime que la media cuando los datos están sesgados), y mantén los ojos abiertos por si en algún caso no te dan un estadístico fundamental (por ejemplo, cuando te dicen la media pero no la desviación estándar).

Capítulo 6

Representación gráfica de datos categóricos

En este capítulo

- Representar visualmente datos categóricos
- Interpretar y valorar diagramas y gráficos

Las presentaciones visuales de datos, en particular los diagramas y los gráficos, están por todas partes y se utilizan para presentar información de todo tipo de cosas, desde los resultados de unas elecciones (desglosados hasta la característica más nimia) hasta la evolución de un índice bursátil a lo largo de los últimos años (meses, semanas, días, minutos). Vivimos en una sociedad donde prima la gratificación instantánea y la información rápida; todo el mundo quiere llegar en seguida al meollo de las cosas y que le ahorren los detalles.

La abundancia de diagramas y gráficos no es necesariamente algo malo, pero hay que andarse con ojo: algunos son incorrectos o incluso inducen a error (a veces queriendo y otras veces sin querer), y es importante saber lo que debes buscar.

Este capítulo trata de los gráficos hechos con *datos categóricos* (datos que sitúan a las personas en grupos o categorías, por ejemplo el sexo, la opinión o si un enfermo toma medicación todos los días). En las próximas páginas te enseñaré a interpretar estas presentaciones visuales y te daré algunos consejos para valorarlas críticamente y detectar posibles problemas. (***Nota:*** las presentaciones visuales de *datos numéricos*, por ejemplo el peso, la nota de un examen o el número de pastillas que un enfermo se toma cada día, se explican en el capítulo 7.)

Los datos categóricos suelen presentarse visualmente mediante gráficos de sectores y gráficos de barras. En este capítulo encontrarás ejemplos de cada tipo de presentación junto con algunas reflexiones sobre su interpretación y consejos para realizar una valoración crítica.

Coge otro trocito de tarta

Los gráficos de sectores separan los datos categóricos por grupos y muestran el porcentaje de personas o elementos que pertenecen a cada grupo. Como los gráficos de sectores tienen forma de tarta, es muy sencillo comparar las "porciones" que representan a los distintos grupos.

Puesto que cada elemento del estudio pertenece a una categoría (y sólo a una), la suma de todas las porciones de la tarta debería ser el 100% o casi el 100% (puede haber un poco de redondeo). De todos modos, ten los ojos bien abiertos por si alguna vez los porcentajes no cuadran.

Contabilidad de gastos personales

Cuando gastas dinero, ¿en qué te lo gastas? ¿Cuáles son tus tres principales fuentes de gasto? Según una encuesta de consumo realizada en 2008 por la Oficina Estadounidense de Estadísticas Laborales, las seis partidas de gasto más importantes en ese país eran la vivienda (33,9%), el transporte (17,0%), la alimentación (12,8%), los seguros y pensiones (11,1%), la sanidad (5,9%) y el ocio (5,6%). Estas seis categorías representan más del 85% de los gastos de un consumidor medio (aunque los porcentajes exactos cambian cada año, las seis partidas principales son siempre las mismas).

La figura 6-1 resume el gasto de los consumidores estadounidenses en 2008 por medio de un gráfico de sectores. Como puedes ver, la categoría "Otros gastos" es bastante grande (13,7%). Lo que ocurre es que, con tantos otros gastos posibles (incluido este libro), a cada uno le correspondería una porción minúscula de la tarta, con lo que el resultado final sería un auténtico embrollo. En este caso resulta demasiado complicado desglosar aún más la categoría "Otros gastos" (pero en muchos otros casos sí se puede).

Lo ideal es que un gráfico de sectores no tenga demasiadas categorías, porque eso distrae al lector de la idea o ideas principales que el gráfico pretende comunicar. Sin embargo, juntar el resto de las categorías en una sola que pasa a ser una de las más grandes del gráfico hace que los lectores se pregunten qué incluye esa categoría. En lo tocante a gráficos y diagramas, hacer bien las cosas obliga a encontrar un delicado equilibrio.

Los ingresos del juego

En Estados Unidos, los organismos estatales de loterías y apuestas generan muchos ingresos y también devuelven gran parte del dinero recaudado (una parte de los beneficios se destina a premios y otra parte a programas estatales, por ejemplo educación). ¿De dónde proceden los ingresos de estos organismos de loterías y apuestas? La figura 6-2 es un gráfico de sectores que muestra los distintos juegos y el porcentaje de

ingresos que representa cada uno, según un informe reciente del Organismo de Loterías y Apuestas del estado de Ohio (si te fijas bien, verás que los porcentajes no suman 100 debido al pequeño error de redondeo).

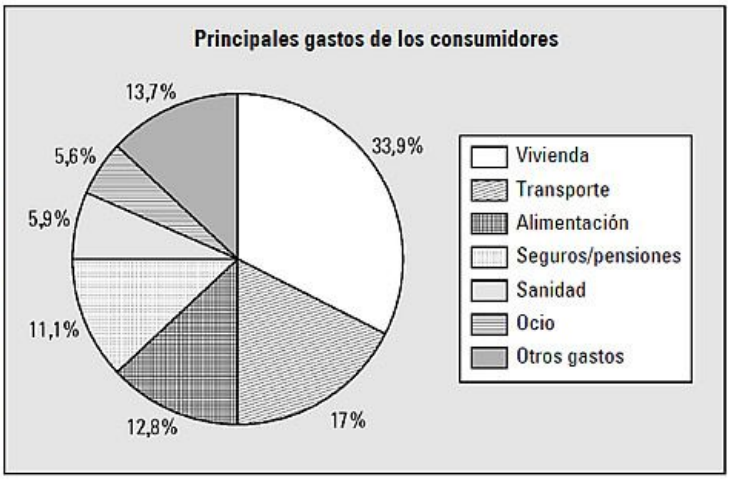


Figura 6-1: Gráfico de sectores que muestra las partidas de gasto de los consumidores estadounidenses

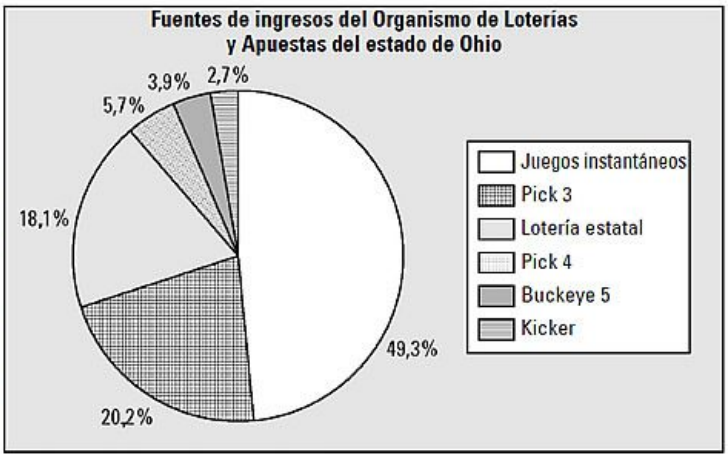


Figura 6-2: Gráfico de sectores donde se desglosan los ingresos de un organismo estatal de loterías y apuestas

Tal y como puedes ver en el gráfico de la figura 6-2, el 49,3% de los ingresos proceden de juegos instantáneos (rasca y gana). El resto procede de diversos juegos de azar en que los jugadores eligen una serie de números y ganan un premio si varios de esos números coinciden con los elegidos en el sorteo. Es el modelo que en España sigue la BonoLoto, por ejemplo.

Este gráfico de sectores no te dice cuánto dinero se recauda, sino tan sólo el porcentaje de dinero que se recauda con cada juego. Aproximadamente la mitad de los ingresos (49,3%) proviene de juegos de rasca y gana, pero no sabemos si esa cantidad asciende a 1 millón de dólares, 2 millones, 10 millones o incluso más. Es imposible saberlo sin conocer la cantidad total de dinero recaudado.

No obstante, consultando otro gráfico en el sitio web del organismo conseguí enterarme de ese dato: la recaudación total (en un período de diez años) fue de "1.983,1 millones de dólares" (unos 1.513 millones de euros). Puesto que el 49,3% de las ventas procedían de juegos instantáneos, esta modalidad representa unos ingresos de 977.668.300 dólares en un período de diez años (745.800 euros). Desde luego, eso es mucho rascar.

Comida para llevar

Cuando miras el gráfico de sectores de una encuesta, también es importante fijarse en los totales. El otro día leía una noticia con los últimos resultados de una encuesta popular. La pregunta era: "¿Qué día de la semana prefieres cenar comida para llevar?". Los resultados se muestran en un gráfico de sectores que tienes en la figura 6-3.

Se ve claramente que la noche de los viernes es la más popular entre los consumidores de comida para llevar (lo cual me parece lógico), y luego la demanda va bajando de sábado a lunes. En realidad, los porcentajes indicados en la figura 6-3 sólo son válidos para las personas encuestadas. La medida en que esos resultados reflejen a la población entera dependerá de muchos factores, uno de los cuales es el tamaño de la muestra. Por desgracia, en el gráfico no se menciona el tamaño de la muestra (por ejemplo, estaría bien que debajo del título hubieran puesto "*n*=XXX"; donde *n* debería ser el tamaño de la muestra).

Sin conocer el tamaño de la muestra, no puedes saber hasta qué punto la información es exacta. ¿Qué resultados te parecen más exactos, los que se basan en 25 personas, 250 personas o 2.500 personas? Cuando te dan el 10%, no sabes si son 10 de cada 100, 100 de cada 1.000 o incluso 1 de cada 10. Para los estadísticos, 1 de cada 10 no es lo mismo que 100 de 1.000, por mucho que en ambos casos se trate del 10% (no se lo digas a un matemático porque te dirá que estás como una cabra).

Los gráficos de sectores generalmente no mencionan el tamaño total de la muestra. Debes comprobar siempre este dato, sobre todo si los resultados son importantes para ti. ¡No des por sentado que es un número muy grande! Si no te dicen el tamaño de la muestra, ponte en

contacto con la fuente de los datos y pídeselo.

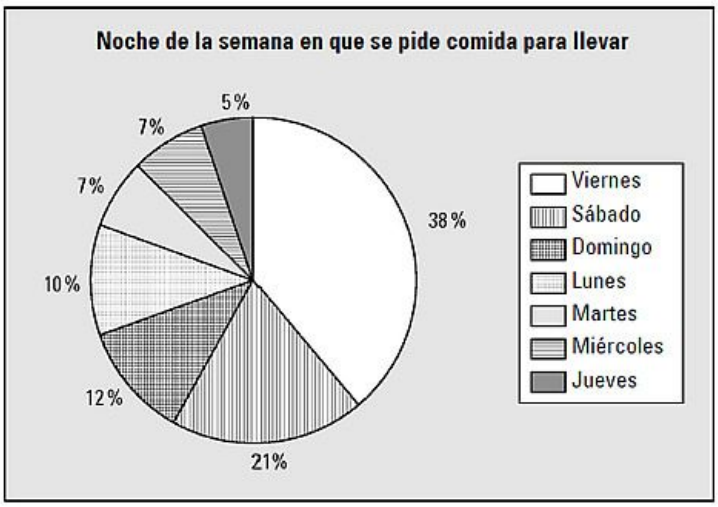


Figura 6-3: Gráfico de sectores con los resultados de una encuesta sobre comida para llevar

Proyección de tendencias de edad

La Oficina del Censo de Estados Unidos proporciona una cantidad ingente de datos, estadísticas y gráficos sobre la población de ese país, incluidas proyecciones de futuro. A menudo se establecen comparaciones entre años al objeto de encontrar variaciones y tendencias.

Un informe reciente de la Oficina del Censo se refería a lo que se ha dado en llamar "las personas mayores" (según la definición del gobierno de Estados Unidos, las personas que tienen sesenta y cinco años o más). Las edades se dividieron en los siguientes grupos: 65-69 años, 70-74 años, 75-79 años, 80-84 años, y 85 años y más. La Oficina calculó y comunicó el porcentaje representado por cada grupo de edad en 2010 y realizó sendas proyecciones para el año 2050.

Yo elaboré gráficos de sectores para los años 2010 y 2050 (proyecciones) con fines de comparación. Puedes ver los resultados en la figura 6-4. Un gráfico de sectores muestra el porcentaje correspondiente a cada grupo de edad en el año 2010, y al lado hay otro gráfico con los porcentajes previstos para 2050 (proyecciones a partir de la edad actual de la población entera de Estados Unidos, las tasas de natalidad y mortalidad, y otras variables).

Si comparas los tamaños de los sectores entre los dos gráficos de la figura 6-4, verás que los sectores de los grupos de edad correspondientes son más grandes en las proyecciones de 2050 (en comparación con 2010) cuanto más viejas son las personas, mientras que los sectores son más pequeños en las proyecciones de 2050 (en comparación con 2010) cuanto menos viejas son las personas. Por ejemplo, el grupo de 65-69 años pasa del 30% en 2010 al 25% previsto para 2050 (es decir, disminuye); mientras que el grupo de 85 años o más pasa del 14% en 2010 al 19% previsto para 2050 (es decir, aumenta).

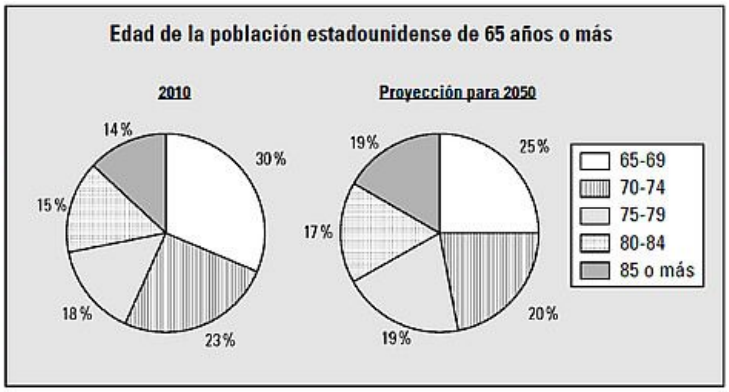


Figura 6-4: Gráficos de sectores sobre la población de edad avanzada, datos de 2010 frente a las proyecciones para 2050

Los resultados de la figura 6-4 indican que la población estadounidense de personas mayores va a envejecer todavía más. A partir de estos datos, los médicos y los investigadores sociales pueden analizar las implicaciones en cuanto a sanidad, cuidado de ancianos, seguridad social, etc.

La clave aquí está en saber si la tendencia va a continuar. Como sabes, el tamaño de la población depende de muchas variables que debes tener en cuenta a la hora de interpretar estas proyecciones de futuro. El gobierno de Estados Unidos siempre incluye advertencias de este tipo en sus informes, lo cual es una prueba de diligencia por su parte.

Los gráficos de sectores de la figura 6-4 resultan útiles para comparar grupos porque están uno al lado del otro, utilizan la misma codificación para las distintas categorías y además los sectores están en el mismo orden (si estuvieran todos desordenados en cada gráfico, tendrías que buscar un grupo de edad determinado en cada gráfico por separado).

Evaluar un gráfico de sectores

Los siguientes consejos te ayudarán a valorar la corrección estadística de un gráfico de sectores:

- ✔ Asegúrate de que la suma de los porcentajes es 100 o casi 100 (el error de redondeo, si lo hay, debe ser muy pequeño).
- ✔ Mucho ojo si el gráfico contiene un sector que lleva por nombre "otros" y es más grande que varios de los otros sectores.
- ✔ Mira a ver si te proporcionan el número total de unidades (personas, importes, etc.) para que puedas saber (grosso modo) cómo era de "grande" la tarta antes de que la dividieran en las porciones que estás viendo.
- ✔ Evita los gráficos de sectores tridimensionales porque no representan las proporciones debidamente. Los sectores de la parte delantera parecen más grandes de lo que son en realidad.

Barra libre de gráficos de barras

Los *gráficos de barras* (también llamados *diagramas de barras*) son quizá la forma de presentación visual más utilizada por los medios de comunicación. Igual que un gráfico de sectores, un gráfico de barras divide los datos categóricos en grupos. Sin embargo, a diferencia de un gráfico de sectores, las cantidades se representan utilizando barras de distintas longitudes. Además, mientras que el gráfico de sectores suele indicar las distintas cantidades por medio de porcentajes, el gráfico de barras puede utilizar el número de elementos de cada grupo (la *frecuencia*) o el porcentaje de elementos que hay en cada grupo (la *frecuencia relativa*).

Gasto en transporte

¿Qué parte de sus ingresos gastan los ciudadanos estadounidenses en desplazarse de casa al trabajo y del trabajo a casa? Depende del dinero que ganan. La Oficina de Estadísticas sobre Transportes (es sorprendente que exista tal cosa, ¿verdad?) realizó un extenso estudio sobre el transporte en Estados Unidos, y gran parte de los datos recopilados se presentaron en forma de gráficos de barras como el de la figura 6-5.

Este gráfico en concreto muestra cuánto dinero gastan en transporte las personas pertenecientes a distintos grupos de ingresos familiares. Según parece, a medida que aumentan los ingresos de la unidad familiar, aumenta también el gasto total en transporte. Esto tiene lógica, ya que cuanto más dinero gana uno, más tiene para gastar.

Pero ¿cambiaría el gráfico de barras si, en lugar de referirnos a cantidades totales de dólares, analizáramos el gasto en transporte como porcentaje de los ingresos de la unidad familiar? Las familias del primer grupo ganan menos de 5.000 dólares al año y tienen que gastar 2.500 dólares en transporte. (***Nota:*** pone "2,5," pero como las unidades están en miles de dólares, son 2.500 dólares.)

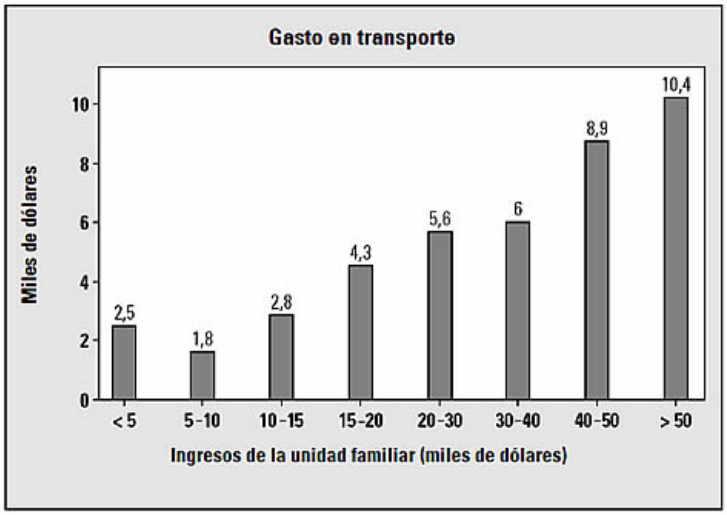


Figura 6-5: Gráfico de barras con los gastos de transporte por grupos de ingresos familiares

Estos 2.500 dólares representan el 50% de los ingresos anuales de las familias que ganan 5.000 dólares al año, y el porcentaje es incluso más alto para quienes ganan menos de esta cantidad. Las familias que ganan entre 30.000 y 40.000 dólares al año gastan 6.000 dólares en transporte, lo que representa entre el 15 y el 20% de los ingresos. Por tanto, aunque las personas que ganan más dinero también gastan más dólares en transporte, este gasto no es mayor si tenemos en cuenta el porcentaje de sus ingresos totales. Según cómo se considere el gasto, el gráfico de barras dará dos lecturas totalmente diferentes.

Por otro lado, también hay que tener en cuenta las categorías utilizadas en el gráfico. Los grupos de ingresos familiares mostrados no son equivalentes. Por ejemplo, las cuatro primeras barras representan ingresos familiares en intervalos de 5.000 dólares, mientras que los tres grupos siguientes muestran incrementos de 10.000 dólares cada uno, y el último grupo incluye a todas las familias que ganan más de 50.000 dólares al año. Los gráficos de barras que utilizan intervalos de distintos tamaños para representar valores numéricos (como ocurre en la figura 6-5) dificultan la labor de comparar los datos (no obstante, estoy segura de que el gobierno tendrá sus razones para presentar las cifras de esta forma; por ejemplo, puede que sea así como desglosan los ingresos en los tramos de los impuestos).

Una última cosa: las categorías de la figura 6-5 se solapan en los límites. Por ejemplo, la cantidad de 30.000 dólares aparece en la quinta y en la sexta barra del gráfico. Entonces, si tus ingresos familiares son de 30.000 dólares, ¿en qué grupo estás? (mirando la figura 6-5 no lo puedes saber, pero seguro que las instrucciones están enterradas en un informe extensísimo que han guardado en el sótano de algún edificio de Washington). Este tipo de solapamientos aparecen con bastante frecuencia en los gráficos, pero es necesario saber qué hacer con los valores límite. Por ejemplo, podría aplicarse la regla siguiente: "Los datos que coincidan exactamente con un valor límite se incluirán en la barra situada inmediatamente a la derecha" (en la figura 6-5, una familia que gane 30.000 dólares deberá estar en la sexta barra, y no en la quinta). Mientras se siga siempre el mismo criterio para todos los límites, no hay ningún problema. La alternativa, que sería describir los límites de ingresos de la quinta barra como "de 20.000 a 29.999,99 dólares", no supone ninguna mejora. En este orden de ideas, los datos sobre ingresos también pueden presentarse con un histograma (lo explico en el capítulo 7), que tiene un aspecto ligeramente distinto.

Los beneficios del juego

Que los organismos de loterías y apuestas ingresan dinero a espuertas es algo de sobras conocido, pero es justo decir que también gastan lo suyo. Entonces, ¿qué beneficios tienen? La figura 6-6 muestra los ingresos y los gastos recientes de cierto organismo estatal de loterías y apuestas de Estados Unidos.

En mi opinión, este gráfico de barras necesita más información para que resulte más comprensible. Las barras de la figura 6-6 no representan conceptos del mismo tipo. La primera barra representa los ingresos, y las demás representan gastos. El gráfico sería mucho más claro si la primera barra no estuviera incluida; por ejemplo, los ingresos totales podrían mencionarse en una nota al pie.

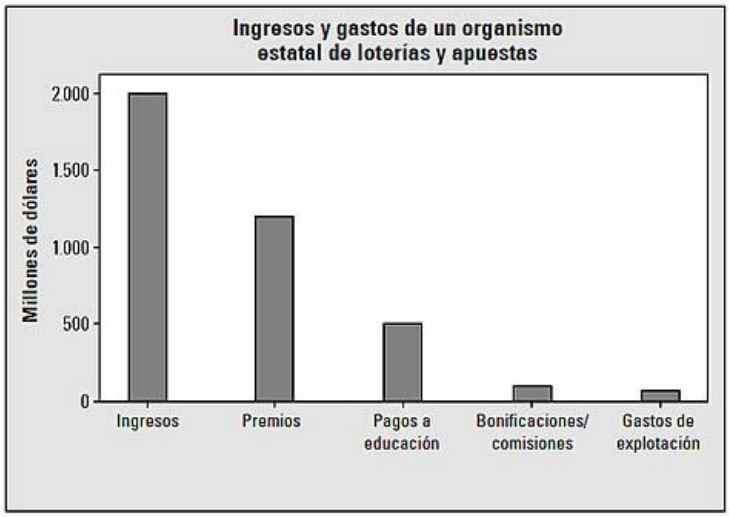


Figura 6-6: Gráfico de barras con los ingresos y los gastos del organismo de loterías y apuestas de cierto estado

Escalas y gráficos de barras

Un gráfico también puede inducir a error según la escala utilizada en el eje de frecuencia/frecuencia relativa (es decir, el eje donde se encuentran las cantidades de cada grupo) o el valor inicial de dicha escala.

Si "estiras" la escala (por ejemplo, que cada centímetro de una barra represente 10 unidades en lugar de 50 unidades), puedes distorsionar la verdad, hacer que las diferencias parezcan mayores o exagerar los valores. Esta distorsión de la verdad también puede ocurrir si el eje de frecuencia comienza en un número que esté muy próximo al punto donde comienzan las diferencias en las alturas de las barras; de este modo básicamente eliminas la parte de abajo de las barras (la menos emocionante) y tan sólo muestras la parte superior, haciendo hincapié (de manera engañosa) en donde está la acción. No todos los ejes de frecuencia tienen que empezar en cero, pero debes estar atento a las situaciones que magnifiquen las diferencias.

Un buen ejemplo de gráfico con la escala estirada aparece en el capítulo 3, en relación con los resultados de los números extraídos en el sorteo Pick 3 (eliges tres números de un solo dígito, y si todos coinciden con los extraídos en el sorteo, te llevas el premio). En el capítulo 3, el porcentaje de veces que se ha extraído cada número (del 0 al 9) se indica en la tabla 3-2, y los resultados se muestran en un gráfico de barras en la figura 3-1a. La escala del gráfico está estirada y comienza en 465, de manera que las diferencias en los resultados parecen mayores de lo que son en realidad; por ejemplo, parece que el número 1 es extraído muchas menos veces, y el número 2 muchas más veces, cuando en realidad no existe diferencia estadística entre el porcentaje de veces que se ha extraído cada número (lo he comprobado).

Entonces, ¿por qué hicieron así el gráfico de la figura 3-1a? La gente podría pensar que tiene información privilegiada y que debería elegir el número 2 porque está en mitad de una buena racha, o el número 1 porque ya le toca subir. Ambas suposiciones son incorrectas, por cierto. Como los números se eligen al azar, lo ocurrido en el pasado no influye en absoluto. La figura 3-1b muestra un gráfico bien hecho. (Para más ejemplos de cómo puede equivocarse la intuición en cuestiones de probabilidad, te recomiendo otro de mis libros, *Probability For Dummies*, también publicado por Wiley, pero no disponible en español.)

Por otro lado, si "acortas" la escala (por ejemplo, que cada centímetro de una barra represente 50 unidades en lugar de 10 unidades), puedes restar importancia a las diferencias y hacer que los resultados parezcan menos significativos de lo que son en realidad. Por ejemplo, puede que un político no quiera llamar la atención sobre un fuerte incremento de la delincuencia a lo largo de su mandato, de manera que, en el gráfico que

muestra el número de delitos de cada tipo, hace que cada barra represente 500 delitos en lugar de 100 delitos. De este modo los números están más juntos y las diferencias son menos apreciables. Su oponente para las próximas elecciones hará justo lo contrario: utilizar una escala estirada para resaltar más el aumento de la delincuencia. (Ahora ya sabes cómo es posible que dos personas manejen los mismos datos y sin embargo lleguen a conclusiones totalmente opuestas. Bienvenido al mundo de la política.)

En el caso de los gráficos de sectores, en cambio, no es posible alterar la escala para poner un énfasis excesivo en los resultados (o para quitarles importancia). Da igual cómo dividas un gráfico de sectores: siempre estarás cortando en porciones una tarta y la proporción correspondiente a cada trozo no cambiará, ni siquiera si dibujas la tarta más grande o más pequeña.

Fastidios cotidianos

En una encuesta reciente se preguntó a 100 oficinistas cuáles eran las cosas que más les molestaban en su lugar de trabajo (antes de continuar leyendo quizá quieras pensar tu respuesta, aunque sólo sea para echarte unas risas). La figura 6-7 muestra un gráfico de barras con los resultados de esa encuesta. Al parecer, la mala gestión del tiempo es lo que más fastidia a esos empleados (espero que no contestaran la encuesta en horario laboral).

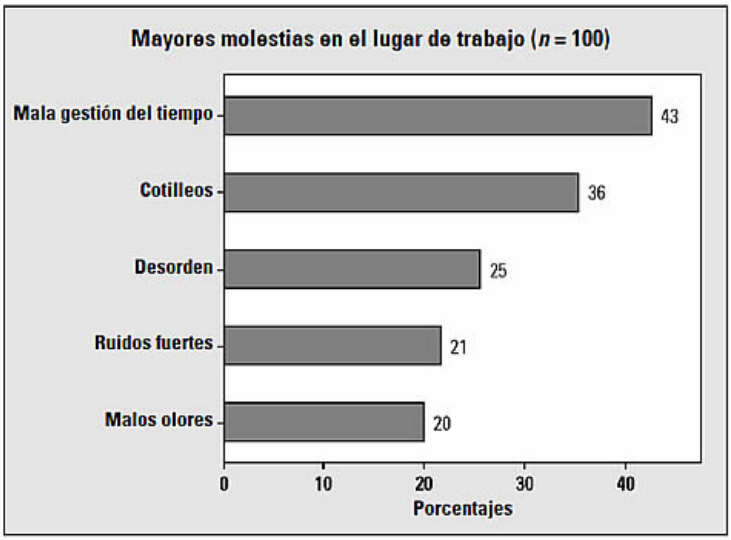


Figura 6-7: Gráfico de barras para los datos de una encuesta con varias respuestas

Evaluar un gráfico de barras

Para hacer las cosas bien, cuando dibujes un gráfico de barras sigue los consejos siguientes:

- ✔ Las barras que dividen valores de una variable numérica (por ejemplo los ingresos) deben ser igual de anchas (si es posible) para que la comparación sea justa.
- ✔ Ten en cuenta la escala del gráfico de barras y asegúrate de que representa la información de forma adecuada.
- ✔ Algunos gráficos de barras no suman uno porque muestran los resultados de más de una variable; deja claro lo que estás resumiendo.
- ✔ Comprueba si los resultados aparecen como porcentaje de elementos en cada grupo (frecuencias relativas) o como cantidad de elementos en cada grupo (frecuencias).
- ✔ Si se trata de frecuencias relativas, comprueba el tamaño total de la muestra: es importante. Si se trata de frecuencias, divide cada una por el tamaño total de la muestra para obtener porcentajes, que son más fáciles de comparar.

Si miras los porcentajes mostrados para cada una de las molestias mencionadas, verás que no suman uno. Por tanto, se deduce que cada una de las personas encuestadas podía mencionar más de una molestia (no es que les costara mucho, seguro); a lo mejor les pidieron que dijeran las tres cosas que más les molestaban, por ejemplo. Con este conjunto de datos, y otros similares que permitan varias respuestas, no es posible utilizar un gráfico de sectores (a menos que hicieras uno para cada una de las molestias indicadas).

Observa que la figura 6-7 es un *gráfico de barras horizontal* (las barras van de lado a lado), mientras que la figura 6-6 es un *gráfico de barras vertical* (las barras están dispuestas verticalmente). Ambas orientaciones son igualmente válidas; puedes utilizar la que más te guste. No obstante, asegúrate de que identificas los ejes correctamente y utilizas las unidades apropiadas (por ejemplo sexo, opinión o día de la semana) en

cada caso.

Capítulo 7

Representación gráfica de datos numéricos

En este capítulo

- Elaborar e interpretar histogramas y diagramas de cajas para datos numéricos
- Analizar los datos numéricos de cronogramas
- Estrategias para detectar gráficos engañosos e incorrectos

El propósito principal de los diagramas y gráficos es resumir unos datos y presentar visualmente los resultados para transmitir un punto de vista de manera clara, eficaz y correcta. En este capítulo hablo de las formas de presentación gráfica utilizadas para resumir datos numéricos, es decir, datos que representan *recuentos* (por ejemplo el número de pastillas que toma diariamente un diabético, o el número de accidentes que ocurren cada año en una determinada intersección) o *mediciones* (el tiempo que tardas en llegar al trabajo o tu tensión arterial).

Verás ejemplos de cómo elaborar, interpretar y evaluar las formas más comunes de representar datos numéricos: cronogramas, histogramas y diagramas de cajas. Asimismo, señalo muchos problemas que pueden surgir con estos gráficos, entre los que están la lectura incorrecta que se hace a menudo. Esta información te ayudará a desarrollar unas habilidades detectivescas importantes para detectar rápidamente los gráficos que induzcan a error.

Historia de un histograma

Un histograma proporciona una impresión visual de todos los datos desglosados en grupos numéricamente ordenados, con lo que obtienes rápidamente una visión general de esos datos, en particular de su forma. En este apartado te enseño a elaborar e interpretar histogramas, y a valorar si son correctos e imparciales.

Elaborar un histograma

Un *histograma* es un gráfico especial que se aplica a datos desglosados en grupos numéricamente ordenados; por ejemplo, grupos de edades como 10-20, 21-30, 31-40, etc. En un histograma las barras están conectadas entre sí, a diferencia de un gráfico de barras (capítulo 6), donde las barras están separadas y representan categorías que no siguen un orden concreto. La altura de cada una de las barras de un histograma representa el número de elementos que hay en su correspondiente grupo *(frecuencia)* o el porcentaje de elementos que hay en cada grupo *(frecuencia relativa)*. Cada elemento del conjunto de datos pertenece a una sola barra.

Puedes hacer un histograma a partir de cualquier conjunto de datos numéricos; sin embargo, no puedes determinar los valores reales del conjunto de datos a partir de un histograma, porque únicamente sabes a qué grupo pertenece cada valor.

Un ejemplo merecedor de un premio

Para todos los amantes del cine que estéis leyendo este libro (en particular los aficionados a las películas antiguas), aquí va un ejemplo de cómo crear un histograma. Los premios Oscar comenzaron a otorgarse en 1928, y una de las categorías más populares ha sido siempre la de mejor actriz en un papel principal. La tabla 7-1 muestra las ganadoras de las ocho primeras estatuillas a la mejor actriz, los años en que ganaron (1928-1935), sus edades en el momento de recibir el premio, y las películas en que participaron. Como puedes ver, las edades van de los 22 a los 62 años, un abanico mucho más amplio de lo que podría uno pensar.

Tabla 7-1. Edades de las ganadoras del Oscar a la mejor actriz 1928-1935

Año	*Ganadora*	*Edad*	*Película*
1928	Laura Gainor	22	*Amanecer*
1929	Mary Pickford	37	*Coqueta*
1930	Norma Shearer	30	*La divorciada*
1931	Marie Dressler	62	*Min and Bill*
1932	Helen Hayes	32	*El pecado de Madelon Claudet*
1933	Katharine Hepburn	26	*Gloria de un día*
1934	Collette Colbert	31	*Sucedió una noche*
1935	Bette Davis	27	*Peligrosa*

Para tener más información sobre las edades de las actrices premiadas, amplié el conjunto de datos al período 1928-2009. La variable "edad" de este conjunto de datos es numérica, de manera que puede representarse gráficamente mediante un histograma. A partir de aquí es posible responder a preguntas como las siguientes: ¿Cuáles son las edades de estas actrices? ¿Son mayoritariamente jóvenes, viejas, de mediana edad? ¿Las edades presentan mucha dispersión o son similares? ¿La mayoría de ellas se encuentran dentro de un intervalo determinado, con unos pocos valores atípicos (actrices muy jóvenes o muy viejas, en comparación con el resto)? Para responder a estas preguntas, la figura 7-1 muestra un histograma de las edades de las actrices galardonadas con un Oscar.

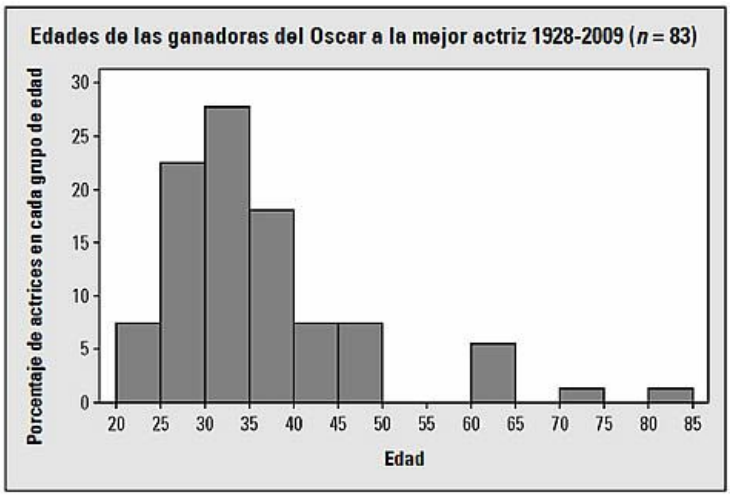


Figura 7-1: Histograma de las edades de las ganadoras del Oscar a la mejor actriz, 1928-2009

Como puedes ver, las edades se indican en el eje horizontal (*x*), en grupos de cinco años: 20-25, 25-30, 30-35 ... 80-85. El porcentaje (frecuencia relativa) de actrices incluidas en cada grupo de edad se indica en el eje vertical (*y*). Por ejemplo, aproximadamente el 27 % de las actrices tenían entre treinta y treinta y cinco años de edad cuando ganaron el Oscar.

Crear grupos apropiados

Para la figura 7-1 utilicé grupos de cinco años porque este incremento determina una periodización natural en la vida de una persona y porque así me salían suficientes barras como para buscar pautas generales. De todos modos, no es imprescindible utilizar este intervalo concreto: a la hora de crear un histograma puedes permitirte alguna que otra licencia poética (sin embargo, esta libertad permite que otros te engañen, como verás más adelante en el apartado "Detectar histogramas engañosos"). Aquí tienes algunos consejos para crear un histograma:

- ✔ Cada conjunto de datos requiere intervalos de diferente tamaño, pero debes evitar los que sean demasiado grandes o demasiado pequeños.
  - Si los intervalos que determinan los grupos de un histograma son muy grandes, todos los datos estarán incluidos en un número de barras muy pequeño, con lo que será imposible realizar comparaciones significativas.
  - Si los intervalos del histograma son muy pequeños, habrá un montón de barras minúsculas que te impedirán obtener una visión general. En ese caso los datos se verán muy revueltos y no revelarán ninguna pauta.
- ✔ Asegúrate de que todos los grupos son igual de grandes. Si una barra es más ancha que las demás, puede que contenga más datos de los que debería, por lo cual, tampoco podrás realizar comparaciones adecuadas

Una idea que podría servirte para elaborar un histograma es coger el rango completo de datos (el más grande menos el más pequeño) y dividirlo por diez para que te salgan diez grupos.

El problema de los valores límite

Siguiendo con el ejemplo de los Oscar, ¿qué pasa si la edad de una actriz está justo en un límite? Por ejemplo, en la tabla 7-1 vemos que en 1930, cuando ganó el Oscar por *La divorciada*, Norma Shearer tenía treinta años. ¿Habría que incluirla en el grupo de edad 25-30 (la barra inferior) o en el grupo de edad 30-35 (la barra superior)?

Mientras seas coherente con todos los puntos de datos, puedes asignar los puntos límite a las barras inferiores respectivas o a las barras superiores respectivas. Lo importante es elegir un criterio y mantener la coherencia. En este caso decidí seguir la norma de poner los valores límite en las barras superiores respectivas, de manera que la edad de Norma Shearer queda dentro de la tercera barra, el grupo de edad 30-35 de la figura 7-1.

Aclaraciones sobre los ejes

La parte más compleja de interpretar un histograma consiste en saber qué se representa en los ejes *x* e *y*. En este sentido, resulta útil emplear buenas etiquetas descriptivas. La mayoría de los programas informáticos etiquetan el eje *x* con el nombre de variable que pusieras al introducir los datos (por ejemplo "edad" o "peso"). Sin embargo, la descripción del eje *y* no está tan clara. De forma predeterminada, muchos programas informáticos etiquetan el eje *y* de un histograma con las palabras "frecuencia" o "porcentaje". Estos términos pueden dar lugar a confusión: ¿frecuencia o porcentaje de qué?

Para que la etiqueta del eje *y* del histograma sea más clara, cambia "frecuencia" por "número de" y añade el nombre de la variable. Si en la etiqueta pone simplemente "porcentaje", es mejor que escribas "porcentaje de" seguido de la variable que corresponda. Por ejemplo, en el histograma de la figura 7-1, en el que se representan las edades de varias actrices oscarizadas, como etiqueta del eje *y* puse "Porcentaje de actrices en cada grupo de edad". En el apartado siguiente te explico cómo interpretar los resultados a partir de un histograma.

Interpretar un histograma

Un histograma te muestra tres características principales de un conjunto de datos numéricos:

- ✔ Cómo están distribuidos los datos entre los grupos (lo que en estadística se llama la *forma* de los datos).
- ✔ El grado de variabilidad de los datos (en términos estadísticos, el grado de *dispersión* de los datos).
- ✔ Dónde está el centro de los datos (para esto los estadísticos utilizan distintas medidas).

Observar la forma de los datos

Una de las características que puedes ver en un histograma es la *forma* de los datos, o dicho de otro modo: la manera en que los datos se reparten entre los grupos. Por ejemplo, podría ocurrir que todos los datos fueran exactamente iguales, en cuyo caso el histograma consistiría en una única barra; o que hubiera el mismo número de datos en todos los grupos, en cuyo caso el histograma sería plano.

Algunos conjuntos de datos tienen una forma característica. En este sentido, hay tres formas que destacan sobre las demás:

- ✔ **Simétrico.** Un histograma es simétrico si al cortarlo por el centro el lado izquierdo es la imagen especular del lado derecho.

  La figura 7-2a muestra un conjunto de datos simétrico que representa el tiempo que tardó en responder a una encuesta cada uno de los 50 encuestados. Como puedes ver, el histograma es casi simétrico.

- ✔ **Sesgado a la derecha.** Un histograma sesgado a la derecha se parece a un montículo ladeado con una cola que se extiende hacia la derecha.

  La figura 7-1, que representa las edades de las actrices oscarizadas, está sesgado a la derecha. En el lado derecho puedes ver unas pocas actrices que son mayores que las demás.

- ✔ **Sesgado a la izquierda.** Si un histograma está sesgado a la izquierda, se parece a un montículo ladeado con una cola que se extiende hacia la izquierda.

  La figura 7-2b muestra un histograma de 17 notas de examen. La forma está sesgada a la izquierda; hay unos pocos estudiantes que sacaron peor nota que todos los demás.

A continuación, algunos consejos para clasificar la forma de un conjunto de datos:

- ✔ **No esperes que unos datos simétricos den un histograma de forma totalmente exacta.** Es muy improbable que los datos se ajusten perfectamente a una determinada pauta, de manera que te corresponde a ti decidir si la forma observada se acerca lo suficiente como para decir que es simétrica.

Si la forma se aproxima tanto a la simetría que otra persona también podría notarlo, y las diferencias son poco importantes, yo la calificaría de simétrica o aproximadamente simétrica. De lo contrario, los datos se consideran no simétricos. (Existen procedimientos estadísticos más sofisticados que comprueban la simetría de los datos, pero están fuera del alcance de este libro.)

✔ **No des por sentado que los datos están sesgados porque el histograma no tenga forma simétrica.** Hay conjuntos de datos de todas las formas y tamaños, y muchos ni siquiera tienen una forma concreta. Incluyo el sesgo en esta lista porque es una de las formas no simétricas más comunes, y una de las que suelen incluirse en los cursos de introducción a la estadística.

Si un conjunto de datos resulta estar sesgado (o eso es lo que interpretas), asegúrate de mencionar la dirección del sesgo (a la izquierda o a la derecha).

Tal y como has visto en la figura 7-1, las edades de las actrices están sesgadas a la derecha. La mayoría de las actrices tenían entre veinte y cincuenta años cuando ganaron el Oscar, y el 27% de ellas estaba en el grupo de edad 30-35. Unas pocas actrices eran más viejas cuando ganaron sus Oscar; aproximadamente el 6% tenía entre sesenta y sesenta y cinco años, y menos del 4% (en total) tenía setenta años o más (si sumas los porcentajes de las dos últimas barras del histograma). Las tres últimas barras son las responsables de que los datos estén sesgados a la derecha.

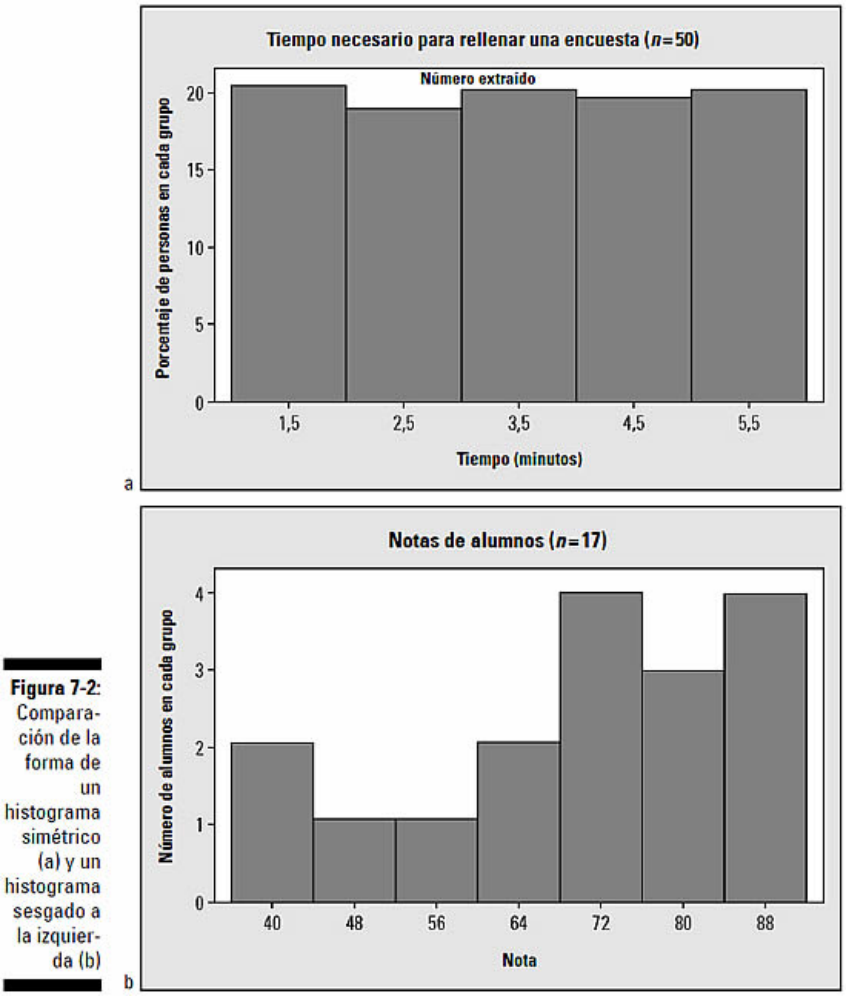


Figura 7-2: Comparación de la forma de un histograma simétrico (a) y un histograma sesgado a la izquierda (b)

Medidas de tendencia central: media y mediana

Un histograma te da una idea aproximada de dónde está el "centro" de los datos. La palabra *centro* está entrecomillada porque existen muchos estadísticos distintos para medir el centro. Las dos medidas de tendencia central más comunes son la media y la mediana. (En el capítulo 5 encontrarás más información sobre las medidas de tendencia central.)

Para visualizar la edad media, piensa que los datos son personas sentadas en un balancín. Tu objetivo es encontrar el punto de equilibrio. Como los datos no cambian de lugar, puedes asumir que las personas permanecen inmóviles y tú desplazas el punto de apoyo (también llamado *fulcro*) hacia el lado que quieras. La media es el lugar donde debes situar el punto de apoyo para que el peso de los dos lados del balancín esté equilibrado.

El punto de equilibrio del balancín depende del peso de las personas que haya en cada lado, no del número de esas personas. Por lo tanto, la media depende de los valores de los datos, y no de la cantidad de datos.

La mediana es el lugar donde pones el punto de apoyo para tener igual número de personas en cada lado del balancín, con independencia de lo que pesen. Al haber el mismo número de personas en cada lado, el balancín no estaría equilibrado en cuanto a peso, a menos que coincidiera el

peso total de las personas de cada lado. Por lo tanto, la mediana no depende de los valores de los datos, sino exclusivamente de su ubicación en el conjunto de datos.

La media se ve afectada por los *valores atípicos*, valores que se encuentran en el extremo superior o inferior del conjunto de datos, alejados del resto de los datos. La mediana, que es el número del centro, no se ve afectada por los valores atípicos.

Visualizar la variabilidad: grado de dispersión en torno a la media

El histograma también permite hacerse una idea de la variabilidad de los datos. Por ejemplo, si todos los datos son iguales, estarán todos en la misma barra y no existirá variabilidad. Si en todos los grupos hay el mismo número de datos, el histograma será plano (todas las barras tendrán más o menos la misma altura) y habrá bastante variabilidad.

La idea de que un histograma plano indica cierto grado de variabilidad quizá vaya en contra de tu intuición, y de ser así no eres el único. Si crees que un histograma plano significa que no existe variabilidad, probablemente estés pensando en un cronograma, donde se representan números sueltos a lo largo del tiempo. (Lee el apartado "Tiempo de cronogramas" más adelante en este mismo capítulo.) De todos modos, piensa que un histograma no representa datos a lo largo del tiempo, sino todos los datos existentes en un momento determinado.

Otra cosa que también causa confusión es la idea de que un histograma con una montaña alta en el centro y colas muy inclinadas por ambos lados tiene menos variabilidad que un histograma plano. Las curvas que parecen montañas en un histograma representan grupos de datos que están muy juntos, mientras que en un histograma plano todos los datos presentan el mismo grado de dispersión, con más variabilidad.

La variabilidad en un histograma es mayor cuando las barras más altas están alejadas de la media, y menor cuando las barras más altas están cerca de la media.

Si miras la figura 7-1, verás que muchas de las actrices oscarizadas están en el intervalo de edad 30-35, y la gran mayoría tienen entre veinte y cincuenta años, lo cual implica bastante diversidad. Luego están los valores atípicos, las pocas actrices de más edad (he contado siete) que dispersan aún más los datos, y que hacen aumentar la variabilidad global.

El estadístico más común para medir la variabilidad en un conjunto de datos es la *desviación estándar*, que, a grandes rasgos, mide el promedio de la distancia que hay entre los datos y la media. La desviación estándar para los datos sobre las edades de las actrices oscarizadas es 11,35 años. (En el capítulo 5 encontrarás más información sobre la desviación estándar.) Una desviación estándar de 11,35 años es bastante grande en el contexto de este problema, pero la desviación estándar se basa en el promedio de distancia respecto de la media, y la media se ve influida por los valores atípicos, de manera que la desviación estándar también. (Hay más información en el capítulo 5.)

Más adelante, en el aparatado "Interpretar un diagrama de cajas", te hablaré de otra medida de variabilidad llamada *rango intercuartílico (RIC)*, que resulta más adecuada cuando tienes datos sesgados.

Relacionar números y gráficos

No puedes calcular medidas de tendencia central y de variabilidad a partir del propio histograma porque no conoces los valores exactos de los datos. Para añadir detalle a tus resultados, además de crear el histograma debes calcular siempre los estadísticos básicos de tendencia central y de variación (todos los estadísticos descriptivos que necesitas, y algunos más, aparecen en el capítulo 5).

La figura 7-1 es un histograma para las edades de las actrices oscarizadas. Como puedes ver, está sesgado a la derecha. Luego, para la figura 7-3, calculé algunos estadísticos (descriptivos) básicos a partir del conjunto de datos. Si haces las cuentas, comprobarás que la edad mediana es 33 años y la edad media es 35,69 años.

La edad media es más alta que la edad mediana debido a que unas pocas actrices eran algo mayores que las demás cuando fueron premiadas. Por ejemplo, Jessica Tandy ganó el Oscar por su papel en *Paseando a Miss Daisy* cuando tenía ochenta y un años, y Katharine Hepburn fue galardonada por *En el estanque dorado* cuando tenía setenta y cuatro. La relación entre la mediana y la media confirma el sesgo (a la derecha) observado en la figura 7-1.

Figura 7-3: Estadísticos descriptivos para las edades de actrices oscarizadas (1928-2009)

Estadísticos descriptivos: edad

Variable	*Recuento total*	*Media*	*Desv est*	*Mínimo*	*Q1*	*Mediana*	*Q3*	*Máximo*	*RIC*
Edad	83	35,69	11,35	21	28	33	39	81	11

Aquí tienes algunos consejos para relacionar la forma del histograma (según lo explicado en el apartado anterior) con la media y la mediana:

- ✔ **Si el histograma está sesgado a la derecha, la media es mayor que la mediana.**

  Esto es así porque los datos sesgados a la derecha tienen unos pocos valores altos que suben la media pero no influyen en dónde se encuentra el centro exacto de los datos (es decir, la mediana). Si miras el histograma de las edades de las actrices oscarizadas, en la figura 7-1, verás que está sesgado a la derecha.

- ✔ **Si el histograma es casi simétrico, la media y la mediana están muy cerca una de otra.**

  "Casi simétrico" significa que los dos lados son muy parecidos; no hace falta que sean exactamente iguales. El significado de "casi" y "cerca" depende del contexto de los datos; por ejemplo, diremos que los números 50 y 55 están cerca si todos los valores están comprendidos entre 0 y 1.000, pero consideraremos que están más alejados si todos los valores se encuentran entre 49 y 56.

El histograma mostrado en la figura 7-2a es casi simétrico. Tanto la media como la mediana son 3,5.

- ✔ **Si el histograma está sesgado a la izquierda, la media es menor que la mediana.**

  Esto es así porque los datos sesgados a la izquierda tienen unos pocos valores pequeños que bajan la media pero no influyen en dónde se encuentra el centro exacto de los datos (es decir, la mediana).

La figura 7-2b representa las notas de examen de 17 alumnos, y los datos están sesgados a la izquierda. Según mis cálculos, la media y la mediana del conjunto de datos original son 70,41 y 74 respectivamente. La media es menor que la mediana debido a que unos pocos alumnos sacaron una nota bastante más baja que los demás. Estos resultados concuerdan con la forma general del histograma mostrado en la figura 7-2b.

Los consejos anteriores para interpretar histogramas también son aplicables a la inversa. Si por alguna razón te falta el histograma de los datos y sólo tienes la media y la mediana, puedes compararlas entre sí para hacerte una idea aproximada de la forma del conjunto de datos.

- ✔ Si la media es mucho más grande que la mediana, generalmente los datos estarán sesgados a la derecha, y habrá unos pocos valores más altos que los demás.
- ✔ Si la media es mucho más pequeña que la mediana, generalmente los datos estarán sesgados a la izquierda, y habrá unos pocos valores pequeños que bajarán la media.
- ✔ Si la media y la mediana están muy cerca una de la otra, sabes que los datos estarán bastante equilibrados en ambos lados, es decir, que serán casi simétricos.

Bajo ciertas condiciones, puedes juntar la media y la desviación estándar para describir un conjunto de datos con bastante detalle. Si los datos presentan una distribución normal (una campana de Gauss con pendientes iguales por ambos lados; está explicada en el capítulo 5), puedes aplicar la regla empírica.

La regla empírica (también descrita en el capítulo 5) dice que si los datos tienen una distribución normal, aproximadamente el 68% de los valores se encuentran a no más de una desviación estándar de la media, aproximadamente el 95% de los valores se encuentran a no más de dos desviaciones estándares de la media, y el 99,7% de los valores se encuentran a no más de tres desviaciones estándares de la media. Estos porcentajes son válidos para la distribución normal (datos con forma de campana) exclusivamente y no pueden aplicarse a conjuntos de datos que tengan otras formas.

Detectar histogramas engañosos

No existen normas fijas para crear un histograma; la persona que elabora el gráfico es quien elige los grupos del eje *x* y también la escala y los puntos inicial y final del eje *y*. Sin embargo, el hecho de que exista este margen discrecional no significa que no pueda uno equivocarse; de hecho, un histograma puede inducir a error por muchos motivos. En los siguientes apartados verás ejemplos de histogramas engañosos y cómo detectarlos.

El problema de que haya pocos grupos

Aunque el número de grupos utilizados está a la entera discreción de la persona que elabora el histograma, es posible meter la pata por poner muy pocas barras, en cuyo caso los datos están todos amontonados, o por poner demasiadas barras, en cuyo caso se magnifican hasta las diferencias más pequeñas.

Para decidir cuántas barras debería haber en un histograma, yo me fijo en los grupos utilizados para formar las barras del eje *x*. Por ejemplo, si un examen se califica sobre 100 puntos, no tiene sentido dividir las notas en grupos de dos puntos (demasiado detalle y, por tanto, demasiadas barras). De igual modo, no tiene sentido agrupar las edades de las actrices en intervalos de veinte años, porque ese criterio no sería suficientemente descriptivo.

Las figuras 7-4 y 7-5 ilustran lo que quiero decir. Cada histograma resume n=222 observaciones del intervalo de tiempo transcurrido entre erupciones del géiser Old Faithful, en el parque de Yellowstone. La figura 7-4 utiliza seis barras que agrupan los datos en segmentos de diez minutos. Este histograma muestra cierto sesgo a la izquierda, pero con 222 observaciones tienes que embutir un montón de datos en sólo seis grupos; por ejemplo, la barra para 75-85 minutos tiene más de 90 elementos de datos. Por tanto, podemos desglosarlo un poco más.

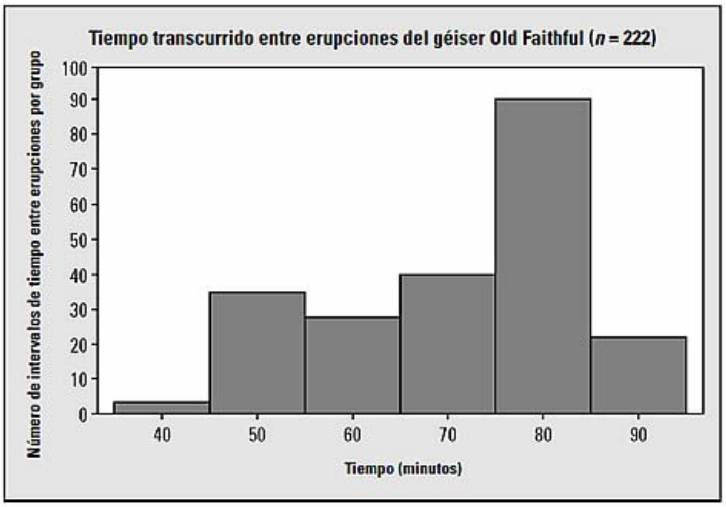


Figura 7-4: Histograma n.° 1, que muestra el tiempo transcurrido entre erupciones del géiser Old Faithful (n = 222)

La figura 7-5 es un histograma del mismo conjunto de datos, pero en este caso el tiempo entre erupciones se clasifica en grupos de tres minutos, con lo que resultan 19 barras. En este histograma los datos revelan una pauta clara que no era visible en la figura 7-4. Se observan dos picos en los datos, uno cerca de la marca de cincuenta minutos y el otro cerca de la marca de setenta y cinco minutos. Un conjunto de datos con dos picos se llama *bimodal*; la figura 7-5 es un ejemplo claro.

Mirando la figura 7-5 puedes concluir que el géiser tiene dos categorías de erupciones; un grupo donde el tiempo de espera entre erupciones es más corto, y otro grupo donde es más largo. Dentro de cada grupo vemos que los datos están bastante cerca del lugar donde se encuentra el pico. Mirando la figura 7-4 no podemos deducir decir tal cosa.

Si el intervalo elegido para los grupos de la variable numérica es muy pequeño, el histograma tendrá demasiadas barras; en ese caso es posible que cueste interpretar los datos porque las alturas de las barras parecerán más variables que lo que deberían. Por el contrario, si los intervalos son muy grandes, habrá muy pocas barras y puede que no detectes alguna pauta interesante en los datos.

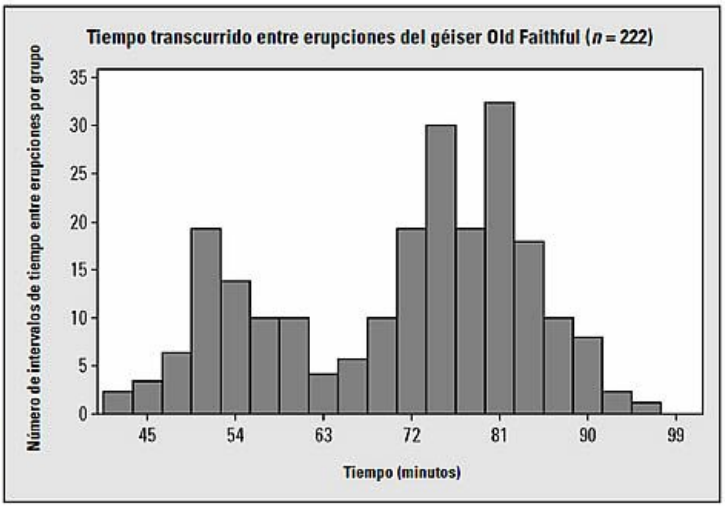


Figura 7-5: Histograma n.° 2, que muestra el tiempo transcurrido entre erupciones del géiser Old Faithful ($n = 222$)

Comprobar la escala y los puntos inicial y final

El eje *y* de un histograma muestra cuántos elementos o individuos hay en cada grupo, para lo cual se utilizan cifras o porcentajes. Un histograma puede inducir a error si la escala y/o los puntos inicial y final del eje *y* no se han elegido bien.

Fíjate bien en la escala del eje *y* de un histograma. Si los incrementos son muy grandes y el punto final es mucho más alto de lo necesario, verás un gran espacio en blanco en la parte superior. Las barras estarán comprimidas, de manera que las diferencias de altura parecerán menores de lo que deberían. Si la escala utiliza incrementos muy pequeños y termina en el valor más pequeño posible, las barras se estiran verticalmente, y entonces las diferencias de altura se exageran y parecen mayores de lo que son en realidad.

Por ejemplo, puedes comparar las escalas de los ejes verticales *(y)* de las figuras 7-5 y 7-6. Cogí los datos del géiser Old Faithful (tiempo transcurrido entre erupciones) y elaboré un histograma con incrementos verticales de veinte minutos, desde 0 hasta 100; mira la figura 7-6. Compara esto con la figura 7-5, donde se utilizan incrementos verticales de cinco minutos, desde 0 hasta 35. La figura 7-6 tiene mucho espacio en blanco y transmite la sensación de que los intervalos de tiempo están distribuidos entre los grupos de manera más uniforme de lo que están en realidad. Además, parece que el conjunto de datos sea más pequeño, a menos que te fijes en lo que pone en el eje *y*. De los dos gráficos, el de la figura 7-5 es más apropiado.

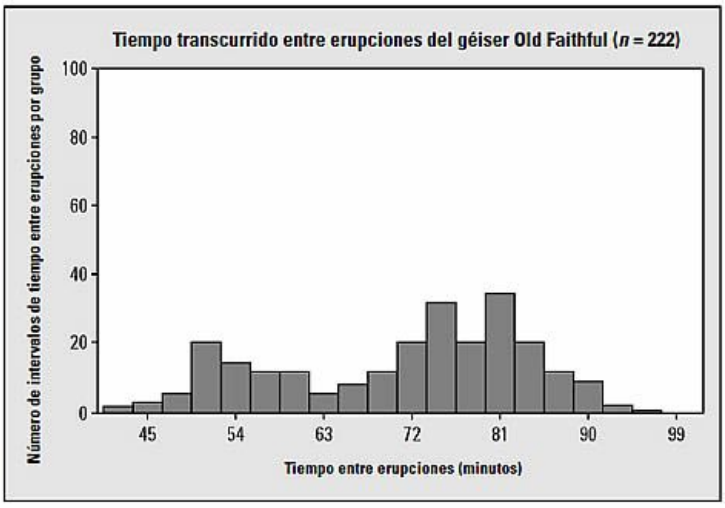


Figura 7-6: Histograma n.° 3, que representa el tiempo transcurrido entre erupciones del géiser Old Faithful

Diagramas de cajas

Un diagrama de cajas es un gráfico unidimensional de datos numéricos basado en el resumen pentanumérico, que incluye el valor mínimo, el percentil 25 (llamado Q_1), la mediana, el percentil 75 (Q_3) y el valor máximo. En esencia, estos cinco estadísticos descriptivos dividen el conjunto de datos en cuatro partes, cada una de las cuales contiene el 25% de los datos. (En el capítulo 5 encontrarás una explicación detallada del resumen pentanumérico.)

Elaborar un diagrama de cajas

Para crear un diagrama de cajas, sigue los pasos siguientes:

1. **Calcula el resumen pentanumérico de tu conjunto de datos (sigue los pasos indicados en el capítulo 5).**

2. **Crea una línea de números vertical (u horizontal) cuya escala incluya los números del resumen pentanumérico y utilice unidades adecuadas que guarden la misma distancia entre sí.**

3. **Marca la posición de cada número del resumen pentanumérico justo encima de la línea de números (para un diagrama de cajas horizontal) o justo a la derecha de la línea de números (para un diagrama de cajas vertical).**

4. **Dibuja una caja delimitada por las marcas correspondientes al percentil 25 y al percentil 75.**

5. **Dentro de la caja, dibuja una línea que pase por la marca correspondiente a la mediana.**

6. **Determina si hay o no valores atípicos.**

Para ello, calcula el RIC (restando Q_3–Q_1) y multiplícalo por 1,5. Añade esa cantidad al valor de Q_3 y réstala al valor de Q_1. De este modo obtienes una frontera más amplia en torno a la media que el tamaño de la caja. Los puntos de datos que queden fuera de esta frontera se consideran valores atípicos.

7. **Si no hay valores atípicos (según los resultados del paso 6), dibuja líneas desde los lados superior e inferior de la caja hasta los valores máximos y mínimos del conjunto de datos.**

8. **Si hay valores atípicos (según los resultados del paso 6), indica su ubicación en el diagrama mediante asteriscos. En lugar de dibujar una línea desde la caja hasta el valor atípico más extremo, interrumpe la línea en el último valor de datos que no sea un valor atípico.**

Muchos paquetes de software (si no la mayoría) señalan los valores atípicos de un conjunto de datos por medio de asteriscos y utilizan el procedimiento descrito en el paso 6 para identificar estos valores atípicos. De todos modos, en algunos casos no se utilizan estos símbolos y procedimientos. Comprueba qué hace tu paquete de *software* antes de analizar unos datos con un diagrama de cajas.

La figura 7-7 muestra un diagrama de cajas horizontal para las edades de las actrices premiadas con un Oscar entre 1928 y 2009. Como puedes ver, los números que separan las secciones del diagrama de cajas coinciden con los estadísticos del resumen pentanumérico indicados en la figura 7-3.

Los diagramas de cajas pueden ser verticales, en cuyo caso los valores del eje van de abajo (el más pequeño) arriba (el más grande), o pueden ser horizontales, en cuyo caso los valores del eje van de izquierda (el más pequeño) a derecha (el más grande). En el siguiente apartado te explico la manera de interpretar un diagrama de cajas.

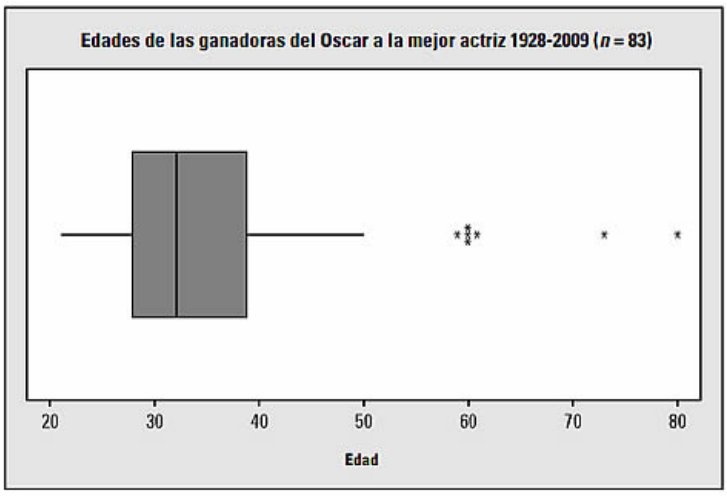


Figura 7-7: Diagrama de cajas para las edades de las actrices premiadas con un Oscar (1928-2009; $n = 83$ actrices)

Interpretar un diagrama de cajas

Igual que un histograma (lee el apartado "Interpretar un histograma"), un diagrama de cajas puede darte información sobre la forma, el centro y la variabilidad de un conjunto de datos. Los diagramas de cajas difieren de los histogramas en cuanto a sus puntos fuertes y débiles, como verás en los siguientes apartados, pero una de sus mayores ventajas es la forma de tratar los datos sesgados.

Observar la forma con cuidado

Un diagrama de cajas puede indicar si un conjunto de datos es simétrico (si lo cortas por el centro, los dos lados son aproximadamente iguales) o está sesgado (inclinado hacia un lado). Si el conjunto de datos es simétrico, la mediana se encuentra más o menos en el centro de la caja. Si los

datos están sesgados, el diagrama de cajas es desigual (la mediana divide la caja en dos partes de distinto tamaño). Si la parte más larga de la caja está a la derecha (o encima) de la mediana, se dice que los datos están *sesgados a la derecha*. Si la parte más larga está a la izquierda (o debajo) de la mediana, los datos están *sesgados a la izquierda*.

Como puedes ver en el diagrama de cajas de la figura 7-7, las edades están sesgadas a la derecha. La parte de la caja que queda a la izquierda de la mediana (la que representa a las actrices más jóvenes) es más corta que la parte de la caja que queda a la derecha de la mediana (la que representa a las actrices de más edad). Esto significa que las edades de las actrices más jóvenes están más cerca unas de otras que las edades de las actrices más viejas. La figura 7-3 muestra los estadísticos descriptivos de los datos y confirma el sesgo a la derecha: la edad mediana (33 años) es menor que la edad media (35,69 años).

Que un lado de la caja sea más largo que el otro no significa que contenga más datos. De hecho, no puedes saber el tamaño de la muestra con sólo mirar un diagrama de cajas (se basa en porcentajes, no en cifras). Pase lo que pase, cada sección del diagrama de cajas (desde el mínimo hasta Q_1, desde Q_1 hasta la mediana, desde la mediana hasta Q_3, y desde Q_3 hasta el máximo) contiene el 25% de los datos. Si una de las secciones es más larga que otra, significa que el rango de los valores de los datos en esa sección es mayor (es decir, los datos están más dispersos). Una sección más pequeña indica que los datos están más condensados (más cerca unos de otros).

Aunque un diagrama de cajas te permite saber si un conjunto de datos es simétrico (cuando la mediana está en el centro de la caja), no puede decirte la forma de la simetría (en cambio, un histograma sí puede). Por ejemplo, la figura 7-8 muestra histogramas de dos conjuntos de datos diferentes, cada uno con 18 valores que varían entre el 1 y el 6. El histograma de la izquierda tiene el mismo número de valores en todos los grupos, mientras que el de la derecha presenta dos picos en el 2 y en el 5. Ambos histogramas revelan que los datos son simétricos, pero sus formas son claramente diferentes.

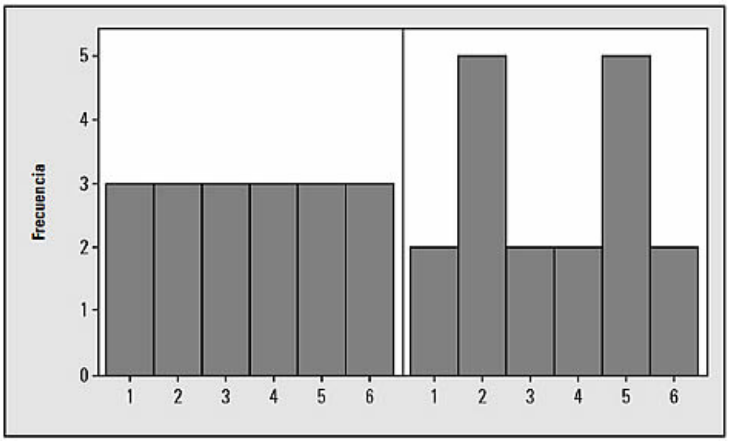


Figura 7-8: Histogramas de dos conjuntos de datos simétricos

La figura 7-9 muestra los diagramas de cajas correspondientes a esos dos conjuntos de datos. Como puedes ver, son exactamente iguales. La razón es que ambos conjuntos de datos tienen los mismos resúmenes pentanuméricos (los dos son simétricos y presentan la misma distancia entre Q_1, la mediana y Q_3). Sin embargo, si sólo vieras los diagramas de cajas y no los histogramas, podrías pensar que los dos conjuntos de datos tienen la misma forma, cuando no es así en absoluto.

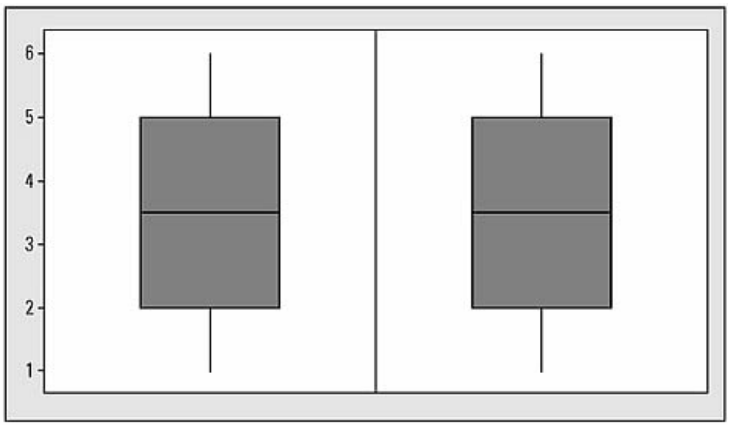


Figura 7-9: Diagramas de cajas de los dos conjuntos de datos simétricos de la figura 7-8

A pesar de que no te dice el tipo de simetría (puedes añadir un histograma a tus análisis para suplir esa carencia), un diagrama de cajas tiene una gran ventaja, y es que te permite identificar directamente las medidas de dispersión y de tendencia central (mientras que un histograma no te lo permite). Además, los diagramas de cajas van bien para comparar conjuntos de datos representándolos en el mismo gráfico, uno al lado del otro.

Todos los gráficos tienen sus ventajas y sus desventajas; por eso siempre es buena idea representar los datos de varias formas distintas.

Medir la variabilidad con el RIC

El rango intercuartílico (RIC) mide la variabilidad en un conjunto de datos descrito por el resumen pentanumérico. El RIC es igual a $Q_3–Q_1$, la diferencia entre el percentil 75 y el percentil 25 (la distancia ocupada por el 50% central de los datos). Cuanto mayor es el RIC, más variable es el conjunto de datos.

Si miramos la figura 7-3, la variabilidad en la edad de las actrices premiadas con un Oscar, medida por el RIC, es $Q_3–Q_1=39–28=11$ años. Del grupo de actrices cuyas edades estaban más cerca de la mediana, la mitad tenían edades que no diferían en más de once años entre ellas cuando fueron galardonadas.

Observa que el RIC no tiene en cuenta los datos situados por debajo del percentil 25 ni por encima del percentil 75, los cuales podrían incluir valores atípicos que a su vez podrían incrementar la medida de variabilidad del conjunto de datos entero. Por tanto, si los datos están sesgados, el RIC es una medida de variabilidad más adecuada que la desviación estándar.

Determinar el centro usando la mediana

La mediana, un elemento del resumen pentanumérico, se representa con la línea que divide la caja en el diagrama de cajas, de manera que resulta muy fácil de identificar. La media, sin embargo, no forma parte del diagrama de cajas y no puede determinarse de forma precisa con sólo mirar el diagrama.

El diagrama de cajas no muestra la media porque está basado totalmente en percentiles. Si los datos están sesgados, la mediana es la medida de tendencia central más adecuada. Naturalmente, puedes calcular la media aparte y añadirla a tus resultados. Nunca está de más indicar ambos estadísticos.

Pesquisas sobre el diagrama de cajas de un géiser

La figura 7-10 muestra los estadísticos descriptivos relevantes para el géiser Old Faithful.

Figura 7-10: Estadísticos descriptivos para los datos del géiser Old Faithful

Estadísticos descriptivos: tiempo entre erupciones

Variable	*Recuento total*	*Media*	*Desv est*	*Mínimo*	*Q1*	*Mediana*	*Q3*	*Máximo*	*RIC*
Tiempo entre erupciones	222	71,009	12,799	42	60	75	81	95	21

A partir de este conjunto de datos, puedes predecir que la forma estará un poco sesgada a la izquierda porque la media es más baja que la mediana (aproximadamente cuatro minutos menos). El RIC es $Q_3–Q_1=81–60=21$ minutos, que es la variabilidad global en el tiempo transcurrido entre erupciones (el 50% de las erupciones están separadas entre sí por un lapso de veintiún minutos).

La figura 7-11 muestra un diagrama de cajas vertical para el tiempo transcurrido entre erupciones del géiser Old Faithful. Se confirma que los datos están sesgados a la izquierda porque la parte de abajo de la caja (donde están los valores pequeños) es más larga que la parte superior.

En el diagrama de cajas de la figura 7-11 puedes ver los valores que marcan el resumen pentanumérico y la información mostrada en la figura 7-10, incluido el de 21 minutos para medir la variabilidad. El centro, marcado por la mediana, es 75 minutos. Esta medida de tendencia central es preferible a la media (71 minutos), la cual baja un poco como consecuencia del sesgo a la izquierda de los valores (unos pocos intervalos de tiempo entre erupciones son más cortos que los demás).

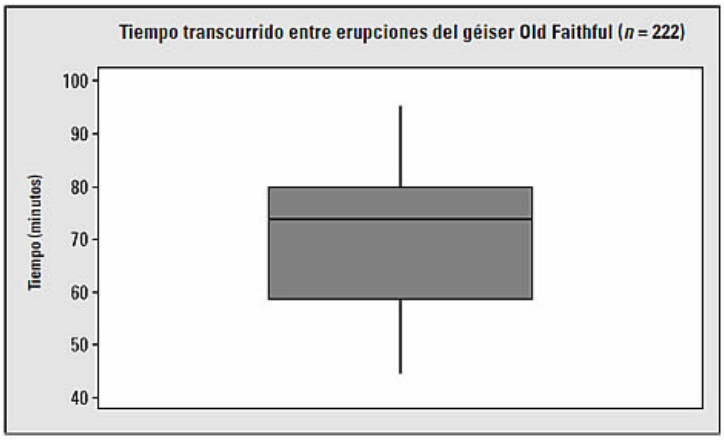


Figura 7-11: Diagrama de cajas con los tiempos transcurridos entre erupciones del géiser Old Faithful ($n = 222$)

Verás que en el diagrama (figura 7-11) no hay valores atípicos señalados con asteriscos. Sin embargo, date cuenta de que el diagrama de cajas no adopta la forma bimodal de los datos que aparece representada en la figura 7-5. Para eso necesitas un buen histograma.

Indicar los valores atípicos

Si miras el diagrama de cajas de la figura 7-7 para las edades de las actrices premiadas con un Oscar, encontrarás varios valores atípicos (siete en total) en el lado derecho del conjunto de datos, identificados con sendos asteriscos (tal y como se indica en el paso 8 del apartado anterior, "Elaborar un diagrama de cajas"). Tres de los asteriscos están apilados porque tres actrices tenían la misma edad, sesenta y un años, cuando fueron galardonadas.

Puedes verificar estos valores atípicos aplicando la regla descrita en el paso 6 del apartado "Elaborar un diagrama de cajas". El RIC es 11 (según se indica en la figura 7-3), de manera que coges $11 \times 1{,}5 = 16{,}5$ años. Al añadir esa cantidad a Q_3 resulta $39 + 16{,}5 = 55{,}5$ años; y al restarla de Q_1 resulta $28 – 16{,}5 = 11{,}5$ años. Por tanto, las actrices que tuvieran menos de 11,5 años (es decir, once años de edad o menos) o más de 55,5 años (es decir, cincuenta y seis años de edad o más) se consideran valores atípicos.

Por supuesto, el extremo inferior de esta frontera (11,5 años) no es relevante porque la actriz más joven galardonada tenía veintiún años (la figura 7-3 muestra que el mínimo es 21). Así pues, sabes que no hay ningún valor atípico en el extremo inferior de este conjunto de datos.

No obstante, sí hay siete valores atípicos en el extremo superior del conjunto de datos, donde se encuentran las actrices de cincuenta y seis años de edad o más. La tabla 7-2 muestra la información de los siete valores atípicos que hay en el conjunto de datos de las edades de actrices premiadas con un Oscar.

Tabla 7-2. Actrices oscarizadas con edades que se consideran valores atípicos

Año	*Nombre*	*Edad*	*Película*
1967	Katharine Hepburn	60	*Adivina quién viene a cenar*
1968	Katharine Hepburn	61	*El león en invierno*
1985	Geraldine Page	61	*Viaje a Bountiful*
2006	Helen Mirren	61	*La reina*
1931	Marie Dressler	62	*Min and Bill*
1981	Katharine Hepburn	74	*En el estanque dorado*
1989	Jessica Tandy	81	*Paseando a Miss Daisy*

El más bajo de los valores atípicos es sesenta años de edad (los que tenía Katharine Hepburn en 1967). Sólo para comparar, la siguiente edad más joven del conjunto de datos es cuarenta y nueve años (Susan Sarandon, 1995). Esto indica un salto claro en este conjunto de datos.

Errores que pueden cometerse al interpretar un diagrama de cajas

Un error común consiste en asociar el tamaño de la caja y la cantidad de datos del conjunto. Recuerda que cada una de las cuatro secciones representadas en el diagrama de cajas contiene el mismo porcentaje (25%) de los datos. El diagrama simplemente marca los lugares que separan esas secciones.

En particular, si la mediana divide la caja en dos partes no iguales, los datos de la parte más grande son más variables que los de la otra parte (en lo que se refiere al rango de los valores). Sin embargo, sigue habiendo la misma cantidad de datos (25%) en la parte más grande y en la parte más pequeña de la caja.

Otro error común tiene que ver con el tamaño de la muestra. Un diagrama de cajas es un gráfico unidimensional con un solo eje que representa la variable que se está midiendo. No hay un segundo eje que te diga cuántos puntos de datos hay en cada grupo. Así pues, si ves dos diagramas de cajas uno al lado del otro, y uno de ellos tiene una caja muy larga y el otro una caja muy pequeña, no deduzcas que la más larga contiene más

datos. La longitud de la caja representa la variabilidad de los datos, no el número de valores de datos.

Cuando analices o crees un diagrama de cajas, asegúrate siempre de que en el título se mencione siempre el tamaño de la muestra *(n)*. Si no te lo dan, no es posible saberlo de otro modo.

Tiempo de cronogramas

Un *cronograma* (también llamado *gráfico de línea temporal*) es una forma de presentación visual utilizada para buscar tendencias en datos distribuidos a lo largo del tiempo (también llamados datos de una serie cronológica). Los cronogramas muestran el tiempo en el eje *x* (por ejemplo por meses, años o días) y los valores de la variable medida en el eje *y* (tasas de natalidad, cifra total de ventas, tamaño de la población, etc.). Cada punto del cronograma resume todos los datos recopilados en ese momento en concreto; por ejemplo, el promedio de todos los precios de la pimienta en el mes de enero o los beneficios totales de 2010.

Interpretar cronogramas

Para interpretar un cronograma, busca pautas y tendencias recorriendo el gráfico de izquierda a derecha.

El cronograma de la figura 7-12 muestra las edades de las actrices premiadas con un Oscar desde 1928 hasta 2009, ordenadas por año. Cada punto indica la edad de una actriz, la que ganara el Oscar ese año. Se observa una pauta cíclica a lo largo del tiempo; es decir, las edades aumentan y disminuyen una y otra vez con cierta regularidad. Es difícil decir qué puede estar pasando: hay muchas variables que intervienen en la elección de la actriz ganadora, por ejemplo el tipo de película, el tipo de papel, el estado de ánimo de los votantes, etc., y algunas de estas variables también podrían seguir una pauta cíclica.

En la figura 7-12 también se observa una levísima pauta alcista, lo cual indicaría que las actrices ganan el Oscar a una edad cada vez mayor. Igual que antes, yo no haría demasiadas suposiciones a partir de este resultado porque los datos presentan un alto grado de variabilidad.

Y ya que hablamos de variabilidad, observa que las edades representadas por los puntos fluctúan bastante en el eje *y* (que representa la edad); todos los puntos están comprendidos entre veinte y ochenta años, y la mayoría entre veinticinco y cuarenta y cinco años, diría yo. Esto concuerda con los estadísticos descriptivos mostrados en la figura 7-3.

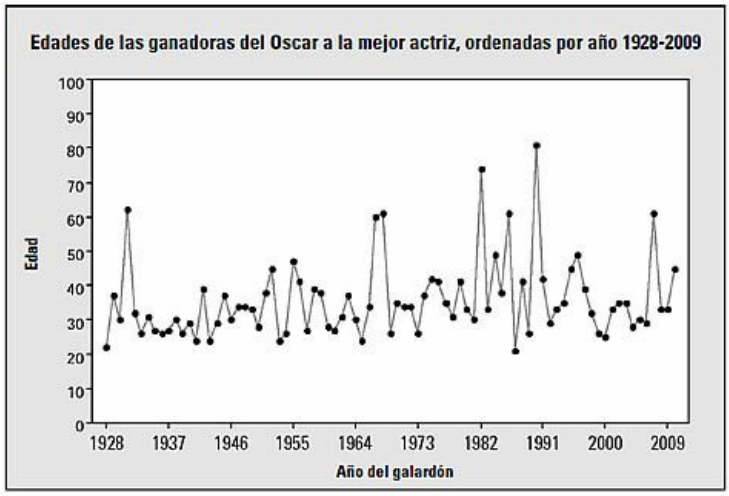


Figura 7-12: Cronograma n.° 1 de las edades de las ganadoras del Oscar a la mejor actriz, 1928-2009

Comprender la variabilidad: cronogramas e histogramas

No hay que confundir la variabilidad en un histograma con la variabilidad en un cronograma. Si los valores cambian con el tiempo, en un cronograma se representan como puntos altos y bajos, y muchos cambios de alto a bajo (a lo largo del tiempo) indican mucha variabilidad. Por tanto, una línea plana en un cronograma indica que no hay cambios ni variabilidad en los valores a lo largo del tiempo. Por ejemplo, si el precio de un producto permanece invariable durante doce meses seguidos, el cronograma correspondiente sería plano.

Pero si todas las barras de un histograma tienen la misma altura, o muy parecida, significa que los datos están uniformemente distribuidos entre todos los grupos, lo cual indica una gran variabilidad en los datos (encontrarás un ejemplo en la figura 7-2a).

Detectar cronogramas engañosos

Igual que con todos los gráficos, tienes que evaluar las unidades de los números representados. Por ejemplo, resultaría engañoso mostrar en un cronograma el número de delitos cometidos, en lugar de la tasa de delincuencia (delitos por habitante). Como el número de habitantes de una ciudad cambia con el tiempo, la tasa de delincuencia es la medida adecuada. Asegúrate de entender qué números se están representando y comprueba si son imparciales y correctos.

Comprobar la escala, el punto inicial y el punto final

La escala del eje vertical influye mucho en el aspecto del cronograma. La figura 7-12 muestra el cronograma original de las edades de las actrices que recibieron un Oscar entre 1928 y 2009, clasificadas en incrementos de diez años. Como hemos comentado antes, se observa bastante variabilidad.

En la figura 7-12, el punto inicial y el punto final del eje vertical son 0 y 100, respectivamente, con lo cual se crea algo más de espacio en blanco en la parte superior e inferior del gráfico. También podría haber utilizado 10 y 90 como puntos inicial y final, pero este gráfico me pareció razonable.

Pero ¿qué pasaría si cambiara el eje vertical? La figura 7-13 muestra los mismos datos pero con el punto inicial en 20 y el punto final en 80. Los incrementos de diez años parecen más largos que los mismos incrementos de diez años de la figura 7-12. Estos dos cambios en el gráfico exageran las diferencias en las edades.

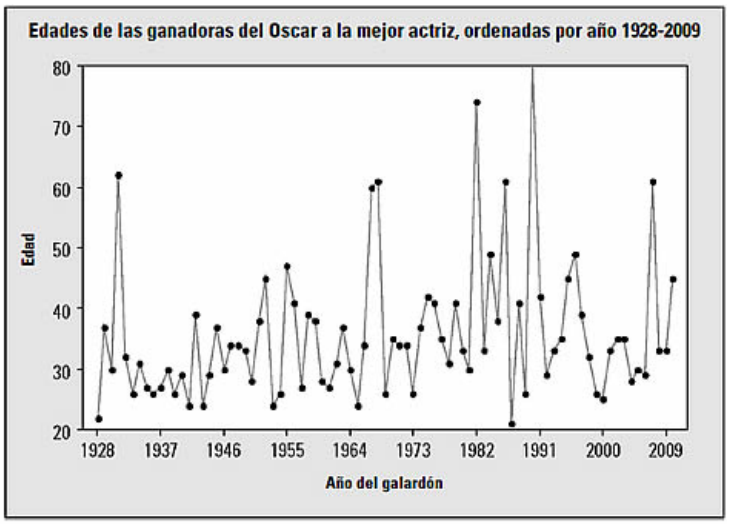


Figura 7-13: Cronograma n.° 2 de las edades de las ganadoras del Oscar a la mejor actriz, 1928-2009

¿Cómo decidir qué gráfico es el idóneo para los datos que tienes? No existe el gráfico perfecto, así que no hay una respuesta correcta o incorrecta, pero sí existen límites. Para detectar rápidamente posibles problemas, fíjate en la escala y en los puntos inicial y final.

Simplificar el exceso de datos

La figura 7-14 muestra un cronograma del tiempo transcurrido entre erupciones del géiser Old Faithful. Este gráfico contiene 222 puntos, cada uno de los cuales representa el tiempo transcurrido entre una erupción y la siguiente, para todas las erupciones ocurridas durante un período de dieciséis días.

Esta figura es muy compleja; está repleta de datos y hay demasiados puntos para distinguir nada. Es como cuando los árboles no te dejan ver el bosque. Como ves, el exceso de información puede ser un problema, sobre todo hoy en día, cuando podemos medir los datos de forma continua utilizando las tecnologías más avanzadas. Me jugaría cualquier cosa a que no tenían a un estudiante de pie junto al géiser anotando en un cuaderno los tiempos transcurridos entre erupciones.

Para tener una imagen más clara de los datos sobre el géiser Old Faithful, combiné todas las observaciones de un único día y calculé la media; repetí esta operación para los dieciséis días y luego representé todas las medias ordenadas en un cronograma. De este modo los puntos de datos pasaron de 222 a 16. El cronograma se muestra en la figura 7-15.

En este cronograma observo una pauta cíclica en los datos: cada día o dos parece que el tiempo entre erupciones aumenta. Aunque estos cambios no son definitivos, sí proporcionan información importante que los científicos pueden tener en cuenta para estudiar el comportamiento de géiseres como el Old Faithful.

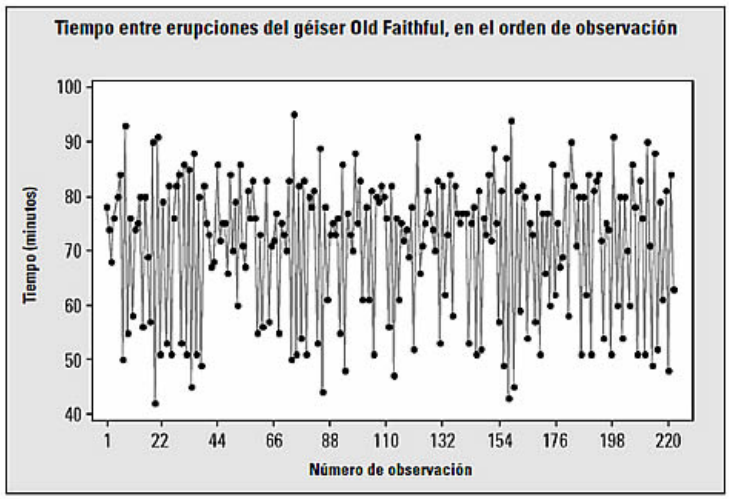


Figura 7-14: Cronograma que muestra el tiempo transcurrido entre erupciones del géiser Old Faithful (n = 222) observaciones consecutivas

Un cronograma condensa en un solo punto todos los datos correspondientes a una unidad de tiempo, mientras que un histograma representa la muestra entera de datos recopilados en esa unidad de tiempo. Por ejemplo, la figura 7-15 muestra el tiempo medio diario transcurrido entre erupciones durante dieciséis días. Para cualquier día, puedes elaborar un histograma de todas las erupciones observadas en ese día en concreto. La combinación perfecta sería mostrar un cronograma de los tiempos medios a lo largo de dieciséis días junto con un histograma que resumiera todas las erupciones de un día en concreto.

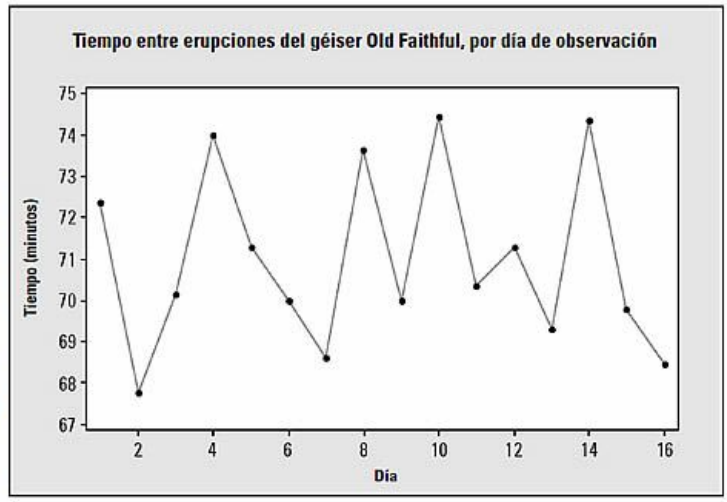


Figura 7-15: Cronograma que muestra el tiempo medio diario transcurrido entre erupciones del géiser Old Faithful (n = 16 días) consecutivas

Evaluar cronogramas

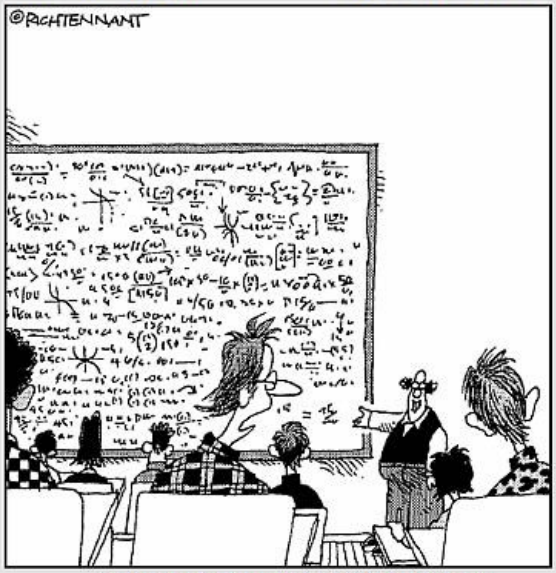
Aquí tienes una lista de control para evaluar cronogramas, con algún comentario añadido:

- ✔ Fíjate en las distancias en la línea temporal de un cronograma. Por ejemplo, resulta engañoso que los puntos del eje horizontal (eje temporal) correspondientes a los años 1990, 2000, 2005 y 2010 aparezcan todos separados por la misma distancia. Esto ocurre cuando los años se tratan como simples etiquetas, en lugar de como números reales.
- ✔ Si tienes una cantidad de datos enorme, plantéate reducirla calculando medias/ medianas para bloques de tiempo y mostrando en el gráfico esas medias/medianas en lugar de todos los datos.
- ✔ Fíjate en la escala, y en los puntos inicial y final del eje vertical (el que muestra los valores de los datos).Si los incrementos son muy grandes o hay mucho espacio en blanco, las diferencias parecen menores de lo que son en realidad, y a la inversa: si los incrementos son pequeños o el gráfico ocupa todo el espacio, las diferencias están exageradas.
- ✔ Como con cualquier otro gráfico, ten muy en cuenta las unidades; asegúrate de que son las adecuadas para realizar comparaciones a lo largo del tiempo. Por ejemplo, ¿las cantidades en unidades monetarias se han ajustado para tener en cuenta la inflación? ¿El gráfico muestra el número de delitos o la tasa de delincuencia?

Parte III

Distribuciones y el teorema del límite central

The 5th Wave Rich Tennant



—¿SE PASA VEINTE MINUTOS ESCRIBIENDO UNA ECUACIÓN Y LUEGO TIENE EL VALOR DE DECIR "YA ESTÁ"?

En esta parte...

Los estadísticos estudian poblaciones; es el pan nuestro de cada día. Hay que medir, contar o clasificar características de una población (utilizando variables aleatorias), encontrar probabilidades y proporciones, y crear (o estimar) resúmenes numéricos para esa población (es decir, parámetros de la población). A veces sabes muchas cosas de una población desde el principio, y otras veces la visión es menos clara. En esta parte consideraremos ambas situaciones.

Si una población se corresponde con una determinada distribución, hay herramientas disponibles para estudiarla. En los capítulos 8, 9 y 10 verás tres distribuciones comúnmente utilizadas: la distribución binomial (para datos categóricos) y la distribución normal y la distribución *t* (para datos numéricos).

Si no conoces los datos concretos de una población (cosa que ocurre continuamente), tomas una muestra y extrapolas los resultados a toda la población. Sin embargo, los resultados de las muestras varían, y eso debes tenerlo en cuenta. En el capítulo 11 investigarás la variabilidad de las muestras, medirás la precisión de los resultados de una muestra y calcularás probabilidades. A partir de ahí podrás estimar parámetros y verificar las afirmaciones realizadas sobre esos parámetros, pero eso ya será en otra parte (la IV, para ser exactos).

Capítulo 8

Variables aleatorias y la distribución binomial

En este capítulo

- Identificar una variable aleatoria binominal
- Encontrar probabilidades utilizando una fórmula o una tabla
- Calcular la media y la variancia

Los científicos y los ingenieros suelen construir modelos de los fenómenos que están estudiando para realizar predicciones y tomar decisiones. Por ejemplo, ¿cuándo y dónde va a tocar tierra este huracán? ¿Cuántos accidentes ocurrirán este año en este cruce si no lo remodelan? ¿Cuál será la población de ciervos en una determinada región dentro de cinco años?

Para contestar a estas preguntas, los científicos (que generalmente colaboran con estadísticos) definen una característica que estén midiendo o contando (por ejemplo el número de cruces, el lugar y el momento en que un huracán toca tierra, el tamaño de la población, etc.) y la tratan como una variable que cambia de forma aleatoria, siguiendo cierta pauta. Estas variables se llaman (lo has adivinado) variables aleatorias. En este capítulo te hablaré sobre las variables aleatorias, sus tipos y características, y por qué son importantes. Y también examinaremos con detalle una de las variables aleatorias más comunes: la binomial.

Definir una variable aleatoria

Una *variable aleatoria* es una característica, una medida o una cifra de recuento que cambia aleatoriamente dentro de un determinado conjunto o con arreglo a una determinada pauta. Su notación es *X*, *Y*, *Z*, etc. En este apartado verás cómo se caracterizan diferentes variables aleatorias y cómo se comportan a largo plazo en lo que se refiere a sus medias y desviaciones estándares.

En matemáticas existen variables como *X* e *Y* que adoptan ciertos valores en función del problema concreto (por ejemplo, determinar la amplitud de un rectángulo), pero en estadística las variables cambian de forma aleatoria. En estadística, el término "aleatorio" significa que no sabes exactamente cuál será el siguiente resultado, pero sí sabes que ciertos resultados ocurren con más frecuencia que otros; la probabilidad de que ocurra algo y de que no ocurra no es la misma (es como cuando yo tiro un rato a canasta; claramente la probabilidad de que enceste no es el 50% y la de que falle no es el 50%. Más bien la probabilidad de acertar es el 5% y la de fallar, el 95%). Esta información puede utilizarse para estudiar mejor datos y poblaciones y para tomar decisiones correctas (por ejemplo, no me pidas que lance yo los tiros libres en tu equipo).

Hay que distinguir entre datos categóricos y datos numéricos (se explica en el capítulo 4). Aunque ambos tipos de datos están asociados a variables aleatorias, yo me referiré exclusivamente a las variables aleatorias numéricas (es lo habitual en la mayoría de los cursos de introducción a la estadística). Para más información sobre análisis de variables categóricas, consulta los capítulos 6 y 19.

Discreto y continuo

Las variables aleatorias numéricas representan cifras y mediciones. Las hay de dos tipos: discretas o continuas, según el tipo de resultados posibles.

- ✔ **Variables aleatorias discretas.** Si los resultados posibles de una variable aleatoria pueden enumerarse utilizando números enteros (por ejemplo 0, 1, 2... 10; o 0, 1, 2, 3), la variable aleatoria es *discreta*.
- ✔ **Variables aleatorias continuas.** Si los resultados posibles de una variable aleatoria únicamente pueden describirse utilizando un intervalo de números reales (por ejemplo, todos los números reales desde 0 hasta infinito), la variable aleatoria es *continua*.

Las variables aleatorias discretas generalmente representan cifras de recuento, por ejemplo el número de personas que votaron sí a la prohibición de fumar entre una muestra aleatoria de 100 personas (los valores posibles son 0, 1, 2 ... 100); o el número de accidentes ocurridos en un determinado cruce a lo largo de un año (los valores posibles son 0, 1, 2...).

Las variables aleatorias discretas pueden ser de dos clases: finitas o infinitas numerables. Una variable aleatoria discreta es *finita* si la lista de valores posibles tiene un número fijo (finito) de elementos (por ejemplo, el número de personas que apoyan la prohibición de fumar en una muestra aleatoria de 100 votantes debe estar comprendido entre 0 y 100). Una variable aleatoria finita muy común es la binomial, que se

comenta con detalle en este mismo capítulo.

Una variable aleatoria discreta es *infinita numerable* si sus valores posibles pueden enumerarse pero no tienen un final concreto. Por ejemplo, el número de accidentes ocurridos en un determinado cruce a lo largo de un período de diez años puede adoptar los valores siguientes: 0, 1, 2... (sabes que la lista termina en algún momento pero no sabes cuándo, así que incluyes todos los números).

Las variables aleatorias continuas generalmente representan mediciones, por ejemplo el tiempo necesario para completar una tarea (un minuto y diez segundos, un minuto y veinte segundos, etc.) o el peso de un recién nacido. Lo que separa las variables aleatorias continuas de las discretas es que las primeras son *infinitas no numerables*: tienen demasiados valores posibles como para enumerarlos o contarlos y/o pueden medirse con un alto grado de precisión (por ejemplo el grado de contaminación atmosférica de una ciudad concreta en un determinado día, medido en partes por millón).

En el capítulo 9 (la distribución normal) y el capítulo 10 (la distribución *t*) encontrarás ejemplos de variables aleatorias continuas comúnmente utilizadas.

Distribuciones de probabilidad

Una variable aleatoria discreta *X* tiene un conjunto definido de valores posibles, cada uno de los cuales tiene una determinada probabilidad de ocurrir. La notación empleada para cualquier resultado específico es una *x* minúscula. Pongamos por caso que lanzas un dado y miras el resultado. La variable aleatoria *X* es el número que salga en el dado (los valores posibles son 1, 2... 6). Si lanzas el dado y te sale un 1, eso es un resultado específico, de manera que escribes "*x*=1".

La probabilidad de que ocurra cualquier resultado específico se designa como *p (x)*, leído "*p* de *x*". De este modo se indica que la probabilidad de la variable aleatoria *X* adopta un valor específico, llamado "*x* minúscula". Por ejemplo, para referirnos a la probabilidad de sacar un 1 en el dado, escribiremos *p* (1).

Los estadísticos utilizan una *X* mayúscula para referirse a variables aleatorias en su forma general; por ejemplo "Supongamos que *X* es el resultado de lanzar un solo dado". La *x* minúscula se utiliza para referirse a resultados específicos de la variable aleatoria, como *x*=1 o *x*=2.

Una lista o función que muestra todos los valores posibles de una variable aleatoria discreta, junto con sus probabilidades, se denomina *distribución de probabilidad*, *p (x)*. Por ejemplo, si lanzas un solo dado, los resultados posibles son 1, 2, 3, 4, 5 y 6, y cada uno de ellos tiene una probabilidad de 1/6 (si el dado no está trucado). Otro ejemplo: imagina que el 40% de los inquilinos de un complejo de apartamentos tiene un perro, el 7% tiene dos perros, el 3% tiene tres perros y el 50% tiene cero perros. Siendo *X*=el número de perros que tiene un inquilino, la tabla 8-1 muestra la distribución de probabilidad para *X*.

Tabla 8-1. Distribución de probabilidad para *X* = número de perros que tiene el inquilino de un apartamento

x	*p(x)*
0	0,50
1	0,40
2	0,07
3	0,03

La media y la variancia de una variable aleatoria discreta

La *media* de una variable aleatoria es el promedio de todos los resultados que cabe esperar a largo plazo (para todas las muestras posibles). Por ejemplo, si lanzas un dado mil millones de veces y anotas los resultados, el promedio de esos resultados es 3,5 (cada resultado tiene la misma probabilidad de salir, de manera que calculas el promedio de los números 1 a 6 y te da 3,5). Sin embargo, si el dado está trucado y el 1 sale más veces que el resto de los números, el resultado promedio de mil millones de tiradas estará más cerca de 1 que de 3,5.

La notación para la media de una variable aleatoria *X* es μ_x o μ (se lee "mu sub *x*", o simplemente "mu *x*"). Puesto que estás considerando los resultados a largo plazo, es lo mismo que buscar la media de una población entera de valores, y por eso se escribe μ_x y no $\bar{x}$ (este símbolo representa la media de una muestra de valores [lee el capítulo 5]). Ponemos la *X* en el subíndice para recordar que la variable a la que pertenece esa media es la variable *X* (a diferencia de una variable *Y* o cualquier otra letra).

La *variancia* de una variable aleatoria es el promedio de los cuadrados de las desviaciones de todos los resultados que obtendrías a largo plazo respecto de la media aritmética de la población, para cada una de las muestras posibles. Esto es lo mismo que la variancia de la población formada por todos los valores posibles. La notación de la variancia de una variable aleatoria *X* es σ_x^2 o σ^2 (se lee "sigma sub *x* al cuadrado" o simplemente "sigma al cuadrado").

La desviación estándar de una variable aleatoria *X* es la raíz cuadrada de la variancia y se designa como σ_X o como σ ("sigma x" o simplemente "sigma"). Más o menos representa el promedio de la desviación respecto de la media.

Igual que ocurre con la media, para designar la variancia y la desviación estándar de una variable aleatoria utilizamos la notación griega. Las notaciones s^2 y s representan la variancia y la desviación estándar de una muestra de elementos, no de la población entera (está explicado en el capítulo 5).

La variancia está en unidades cuadradas, de manera que no puede interpretarse fácilmente. Para la interpretación se utiliza la desviación estándar porque está en las unidades originales de *X*. La desviación estándar puede interpretarse de manera aproximada como el promedio de la desviación respecto de la media.

Identificar una variable binomial

La variable aleatoria discreta más conocida y estimada es la variable binomial. "Binomial", o "binómico", significa "que tiene dos nombres", y se asocia a situaciones en que existen dos resultados; por ejemplo sí/no, o éxito/fracaso (encontrar un semáforo en rojo o no, desarrollar un efecto secundario o no). Este apartado trata sobre la variable aleatoria binomial: cuándo se utiliza, cuál es su probabilidad y cómo se encuentra su media y su variancia.

Una variable aleatoria es binomial (es decir, tiene una distribución binomial) cuando se cumplen las cuatro condiciones siguientes:

1. Hay un número fijo de intentos (*n*).
2. Cada intento tiene dos resultados posibles: éxito o fracaso.
3. La probabilidad de éxito (la llamaremos *p*) es la misma en cada intento.
4. Los intentos son independientes, es decir, el resultado de un intento no influye en el resultado de los demás. Siendo *X* el número total de éxitos en *n* intentos; si se cumplen las cuatro condiciones, *X* tiene una distribución binomial con una probabilidad de éxito (en cada intento) igual a *p*.

La *p* minúscula es la probabilidad de obtener un resultado de éxito en un único intento. No es lo mismo que *p(x)*, que significa la probabilidad de obtener *x* éxitos en *n* intentos.

Comprobar paso a paso las condiciones de distribución binomial

Lanzas una moneda al aire 10 veces y cuentas el número de caras (*X*). ¿*X* tiene una distribución binomial? Para comprobarlo, comprueba las respuestas a las preguntas y afirmaciones de la siguiente lista:

1. ¿Hay un número fijo de intentos?

Lanzas la moneda 10 veces, que es un número fijo. La condición 1 se cumple, y $n=10$.

2. ¿Cada intento tiene solamente dos resultados posibles, éxito o fracaso?

El resultado de cada lanzamiento es cara o cruz, y tú estás interesado en contar el número de caras. Por lo tanto, éxito=cara, y fracaso=cruz. La condición 2 se cumple.

3. ¿La probabilidad de éxito es la misma en cada intento?

Si damos por sentado que la moneda no está trucada, la probabilidad de éxito (que salga cara) es $p=½$ para cada intento. También sabes que $1–½=½$ es la probabilidad de fracaso (que salga cruz) para cada intento. La condición 3 se cumple.

4. ¿Los intentos son independientes?

Partimos de que la moneda se lanza siempre de la misma forma, de manera que el resultado de un lanzamiento no afecta al resultado de los lanzamientos posteriores. La condición 4 se cumple.

Como la variable aleatoria *X* (el número de éxitos [caras] que ocurren en diez intentos [lanzamientos]) cumple las cuatro condiciones, concluimos que tiene una distribución binomial con $n=10$ y $p=½$.

Sin embargo, no todas las situaciones que parecen binomiales lo son en realidad. A continuación pongo algunos ejemplos que ilustran lo que quiero decir.

El número de intentos no es fijo

Supón que vas a lanzar una moneda no trucada hasta que te salgan cuatro caras y vas a contar el número de lanzamientos que tardas en conseguirlo; en este caso *X*=número de lanzamientos. Sin duda, esto parece una situación binomial: la condición 2 se cumple porque tienes éxito

(cara) y fracaso (cruz) en cada lanzamiento; la condición 3 se cumple, ya que la probabilidad de éxito (cara) es la misma (0,5) para cada lanzamiento; y además los lanzamientos son independientes, con lo que se cumple la condición 4.

Sin embargo, *X* no cuenta el número caras, sino el número de intentos necesarios para que salgan cuatro caras en total. Lo que es fijo es el número de éxitos (*X*), y no el número de intentos (*n*). La condición 1 no se cumple, de manera que *X* no tiene una distribución binomial en este caso.

Hay más resultados, aparte de éxito o fracaso

En algunas situaciones que parecen binomiales, en realidad hay más de dos resultados posibles. Por ejemplo, supongamos que lanzas un dado 10 veces y *X* es el resultado de cada lanzamiento (1, 2, 3... 6). Tienes una serie de $n=10$ intentos, dichos intentos son independientes y la probabilidad de cada resultado es la misma en todos los lanzamientos. No obstante, cada vez estás anotando el resultado de lanzar un dado de seis caras, un número del 1 al 6. Ésta no es una situación de éxito/fracaso, de manera que la condición 2 no se cumple.

Sin embargo, según lo que estés anotando, situaciones que originalmente tienen más de dos resultados pueden entrar en la categoría de binomiales. Por ejemplo, si lanzas un dado 10 veces y cada vez anotas si sale o no sale un uno, entonces la condición 2 se cumple porque los dos resultados de interés son que salga un uno ("éxito") y que no salga un uno ("fracaso"). En este caso *p* (la probabilidad de éxito)$=1/6$, y $5/6$ es la probabilidad de fracaso. Por tanto, si *X* es contar el número de unos que te salen en 10 lanzamientos, *X* es una variable aleatoria binomial.

Los intentos no son independientes

La condición de independencia se incumple cuando el resultado de un intento afecta a otro intento. Imagina que quieres conocer la opinión de las personas adultas de tu ciudad acerca de un casino que podría abrirse en el futuro. En lugar de coger una muestra aleatoria de, pongamos, 100 personas, para ahorrar tiempo seleccionas a 50 matrimonios y preguntas a todo el mundo cuál es su opinión. En este caso es razonable decir que los matrimonios tienen una probabilidad mayor de coincidir en sus opiniones que otras personas seleccionadas al azar, de manera que la condición de independencia 4 no se cumple.

La probabilidad de éxito *(p)* varía

Tienes diez personas (seis mujeres y cuatro hombres) y quieres formar un comité de dos personas elegidas al azar. *X* es el número de mujeres en ese comité de dos personas. La probabilidad de elegir una mujer al azar en el primer intento es $6/10$. Como no puedes seleccionar esa mujer otra vez, la probabilidad de seleccionar otra mujer es ahora $5/9$. El valor de *p* ha cambiado, y la condición 3 no se cumple.

Si la población es muy grande (por ejemplo todos los europeos mayores de edad), *p* igualmente cambia cada vez que eliges a alguien, pero este cambio es tan insignificante que no se tiene en cuenta. Igualmente decimos que los intentos son independientes y la probabilidad de éxito, *p*, no varía (así la vida es mucho más sencilla).

Encontrar probabilidades binomiales utilizando una fórmula

Una vez que hayas identificado que *X* tiene una distribución binomial (se cumplen las cuatro condiciones del apartado "Comprobar paso a paso las condiciones de distribución binomial"), posiblemente quieras conocer las probabilidades de *X*. La buena noticia es que no tienes que empezar desde cero: puedes utilizar unas fórmulas establecidas para encontrar probabilidades binomiales utilizando los valores de *n* y *p* propios de cada problema. Las probabilidades de una variable aleatoria binomial *X* pueden encontrarse utilizando la siguiente fórmula para *p(x)*: donde:

$$\binom{n}{x} p^x (1-p)^{n-x}$$

- ✔ *n* es el número fijo de intentos.
- ✔ *x* es el número especificado de éxitos.
- ✔ *n–x* es el número de fracasos.
- ✔ *p* es la probabilidad de éxito en un intento cualquiera.
- ✔ 1–*p* es la probabilidad de fracaso en un intento cualquiera. (***Nota:*** en algunos libros de texto se utiliza la letra *q* para designar la probabilidad de fracaso, en lugar de 1–*p*.) Estas probabilidades son ciertas para cualquier valor de *X* entre 0 (el número más pequeño de éxitos posibles en *n* intentos) y *n* (el número más grande de éxitos posibles).

El número de maneras en que pueden obtenerse *x* éxitos en *n* intentos se llama "*n* sobre *x*", y la notación matemática es $\binom{n}{x}$. Es importante señalar que esta expresión matemática no es una fracción, sino una forma abreviada de representar el número de maneras posibles de ordenar los resultados.

En general, para calcular "*n* sobre *x*" se utiliza la siguiente fórmula:

$$\binom{n}{x}=\frac{n!}{x!(n-x)!}$$

La notación *n! (n factorial* o *factorial de n)* indica el número de maneras de reordenar *n* elementos. Para calcular *n*! hay que multiplicar *n(n–1)(n–2)* ... (2)(1). Por ejemplo, 5! es 5(4)(3)(2)(1)=120; 2! es 2(1)=2; y 1! es 1. Por convención, 0! es igual a 1.

Imagina que tienes que pasar por tres semáforos para llegar a tu lugar de trabajo. Supongamos que *X* es el número de semáforos que encuentras en rojo. ¿Cuántas maneras hay de que encuentres dos semáforos en rojo de camino al trabajo? Bueno, podrías encontrarte en verde el primero y luego los otros dos en rojo; o verde el del medio y rojo el primero y el tercero; o también rojo el primero, rojo el segundo y verde el tercero. Siendo V=verde y R=rojo, puedes escribir estas tres posibilidades de la manera siguiente: VRR, RVR, RRV. Por tanto, hay tres maneras en las que puedes encontrarte dos semáforos en rojo de camino al trabajo, ¿no?

Veamos las operaciones. En este ejemplo, un "intento" es un semáforo, y un "éxito" es encontrarlo en rojo (ya sé que parece raro, pero un éxito es cualquier cosa que estés interesado en contar, sea bueno o malo). Por lo tanto tienes *n*=3 semáforos en total, y estás interesado en la situación en que te encuentras *x*=2 luces en rojo. Utilizando la notación matemática $\binom{3}{2}$, tenemos que "tres sobre dos" es el número de maneras de ordenar dos éxitos en tres intentos.

Para calcular "3 sobre 2," haz lo siguiente:

$$\binom{3}{2}=\frac{3!}{2!(3-2)!}=\frac{3(2)(1)}{[(2)(1)](1)}=\frac{6}{2}=3$$

Esto confirma las tres posibilidades arriba mencionadas para encontrarte con dos semáforos en rojo.

Ahora supongamos que los semáforos son independientes entre sí y que cada uno tiene una probabilidad del 30% de estar en rojo. Imagina que quieres encontrar la distribución de probabilidad para *X*, es decir, una lista de todos los valores posibles de *X* (0, 1, 2, 3) y sus probabilidades.

Antes de ponerte a hacer cálculos, empieza por comprobar las cuatro condiciones (repasa, si lo necesitas, el apartado "Comprobar paso a paso las condiciones de distribución binomial") para ver si realmente se trata de una situación binomial. Aquí *n*=3 intentos (semáforos), compruébalo. Cada intento puede ser un éxito (semáforo en rojo) o un fracaso (semáforo en ámbar o en verde; o dicho de otro modo, semáforo "no rojo"), compruébalo. Los semáforos funcionan de manera independiente, de modo que se cumple la condición de independencia, y como cada luz está encendida el 30% del tiempo, sabes que *p*=0,30 para cada color. Por tanto, *X*=número de semáforos en rojo tiene una distribución binomial. Si vamos a las fórmulas, 1–*p*=probabilidad de un semáforo no rojo=1–0,30=0,70; y el número de semáforos no rojos es 3–*X*.

Aplicando la fórmula de *p(x)* se obtienen las probabilidades para *x*=0, 1, 2 y 3 semáforos en rojo:

$$p(0)=\binom{3}{0}0{,}30^0(1-0{,}30)^{3-0}=$$

$$\frac{3!}{0!(3-0)!}(0{,}30)^0(0{,}70)^3=1(1)(0{,}343)=0{,}343;$$

$$p(1)=\binom{3}{1}0{,}30^1(1-0{,}30)^{3-1}=$$

$$\frac{3!}{1!(3-1)!}(0{,}30)^1(0{,}70)^2=3(0{,}30)(0{,}49)=0{,}441;$$

$$p(2)=\binom{3}{2}0{,}30^2(1-0{,}30)^{3-2}=$$

$$\frac{3!}{2!(3-2)!}(0{,}30)^2(0{,}70)^1=3(0{,}09)(0{,}70)=0{,}189;\ \text{y}$$

$$p(3)=\binom{3}{3}0.30^3(1-0.30)^{3-3}=$$

$$\frac{3!}{3!(3-3)!}(0{,}30)^3(0{,}70)^0=1(0{,}027)(1)=0{,}027.$$

La tabla 8-2 muestra la distribución final de probabilidad para *X*. Observa que todas estas probabilidades suman 1 porque se han incluido y tenido en cuenta todos los valores posibles de *X*.

Tabla 8-2. Distribución de probabilidad para X = número de semáforos en rojo (n=3, p=0,30)

X	$p(x)$
0	0,343
1	0,441
2	0,189
3	0,027

Encontrar probabilidades con la tabla binomial

En el apartado anterior manejábamos valores de *n* bastante pequeños, pero quizá te preguntes cómo te las vas a apañar con la fórmula para calcular probabilidades binomiales cuando *n* sea una cifra alta. Tranquilo, no hay de qué preocuparse. En la tabla binomial que he incluido en el apéndice encontrarás un montón de probabilidades binomiales. Te explico cómo se usa: dentro de la tabla binomial hay varias minitablas, cada una de las cuales corresponde a un valor diferente de *n* (encontrarás *n*=1, 2, 3... 15 y 20). Cada minitabla tiene filas y columnas. Al lado de cada minitabla se encuentran todos los valores posibles de *X* ordenados desde 0 hasta *n*, cada uno con su propia fila. Las columnas de la tabla binomial representan distintos valores de *p* desde 0,10 hasta 0,90.

Encontrar probabilidades para valores concretos de *X*

Para utilizar la tabla binomial del apéndice con el fin de encontrar probabilidades para *X*=número total de éxitos en *n* intentos, siendo *p* la probabilidad de éxito en cualquier intento, sigue los pasos siguientes:

1. **Encuentra la minitabla correspondiente a tu valor de *n* (el número de intentos).**
2. **Encuentra la columna que representa tu valor de *p* (o el valor más próximo al tuyo, si procede).**
3. **Encuentra la fila que representa el número de éxitos *(x)* en el que estás interesado.**
4. **Encuentra la intersección de la fila y la columna de los pasos 2 y 3.** De este modo obtienes la probabilidad de *x* éxitos, expresada como *p (x)*.

Volviendo al ejemplo de los semáforos incluido en el apartado "Encontrar probabilidades binomiales utilizando una fórmula", puedes utilizar la tabla binomial (tabla A-3 del apéndice) para verificar los resultados obtenidos con la fórmula binomial y recogidos en la tabla 8-2. Ve a la minitabla donde n = 3 y mira en la columna donde p=0,3. Encontrarás cuatro probabilidades: 0,343, 0,441, 0,189 y 0,027; son las probabilidades para X=0, 1, 2 y 3 semáforos en rojo, respectivamente, y coinciden con las que aparecen en la tabla 8-2.

Encontrar probabilidades para *X* mayor que un valor, menor que un valor o entre dos valores

La tabla binomial (tabla A-3 del apéndice) muestra probabilidades cuando X es igual a cualquier valor desde 0 hasta n, para distintos valores de p. Para encontrar las probabilidades cuando X es menor o mayor que un valor, o se encuentra entre dos valores, basta con encontrar los valores correspondientes en la tabla y sumar sus probabilidades. En el ejemplo de los semáforos, hay que contar el número de veces (X) que te encuentras un semáforo en rojo (de los tres posibles). Cada semáforo tiene una probabilidad de 0,3 de estar rojo, de manera que tienes una distribución binomial con n = 3 y p = 0,3. Si quieres saber la probabilidad de encontrarte con más de un semáforo en rojo, p (x > 1), tienes que sumar p (2) + p (3) de la tabla A-3, con lo que obtienes 0,189 + 0,027 = 0,216.

La probabilidad de encontrarte entre uno y tres semáforos en rojo es p ($1 \le x \le 3$)=0,441+0,189+0,027=0,657.

Cuando se trabaja con variables aleatorias discretas, hay que distinguir entre una probabilidad *mayor que* (>) y una probabilidad *mayor o igual que* (≥). Volviendo a los dos ejemplos anteriores, puedes ver que p (x>1)=0,216 pero p ($x \ge 1$)=0,657. En el caso de las variables aleatorias continuas, esta distinción desaparece (se explica en el capítulo 9).

Otras frases que debes recordar: "como mínimo" significa ese número o uno mayor, y "como máximo" significa ese número o uno menor. Por ejemplo, la probabilidad de que X sea como mínimo 2 es $p(x \ge 2)$; la probabilidad de que X sea como máximo 2 es $p(x \le 2)$.

Comprobar la media y la desviación estándar de la distribución binomial

Como la distribución binomial se utiliza mucho, los estadísticos se pusieron manos a la obra para dar con fórmulas bonitas y sencillas que

permitieran encontrar su media, su variancia y su desviación estándar (es decir, aplicaron los métodos descritos en el apartado "Definir una variable aleatoria" a las fórmulas de distribución binomial, hicieron todos los cálculos matemáticos y nos presentaron los resultados en bandeja de plata. ¿No te parece genial?). Los siguientes resultados son el fruto de esos esfuerzos.

Si X tiene una distribución binomial con n intentos y una probabilidad de éxito p en cada intento, entonces:

1. La media de X es $\mu=np$.
2. La variancia de X es $\sigma^2=np(1-p)$.
3. La desviación estándar de X es $\sigma=\sqrt{np(1-p)}$

Por ejemplo, pongamos que lanzas una moneda al aire 100 veces y X es el número de caras. En ese caso X tiene una distribución binomial en la que $n=100$ y $p=0{,}5$. La media es $\mu=np=100\ (0{,}50)$ caras (lo cual tiene sentido, ya que la probabilidad de que salga cara y de que salga cruz es la misma). La variancia de X es $\sigma^2=np\ (1-p)=100\ (0{,}5)\ (1-0{,}5)=25$, que está en unidades cuadradas (de manera que no lo puedes interpretar), y la desviación estándar es la raíz cuadrada de la variancia, o sea, 5. Esto significa que si lanzas una moneda al aire 100 veces, y haces eso repetidamente, el número medio de caras que te saldrán es 50, y cabe esperar que esta cifra varíe 5 caras en promedio.

La fórmula para calcular la media de una distribución binomial tiene un significado intuitivo. La p de la fórmula representa la probabilidad de un éxito, sí, pero también representa la *proporción* de éxitos que cabe esperar en n intentos. Por tanto, el *número* total de éxitos que cabe esperar (es decir, la media de X) es $\mu=np$.

La fórmula de la variancia también tiene un significado intuitivo. La única variabilidad en los resultados de cada intento está en el éxito (con probabilidad p) y el fracaso (con probabilidad $1-p$). A lo largo de n intentos, la variancia del número de éxitos/fracasos se mide con $\sigma^2=np\ (1-p)$. La desviación estándar no es más que la raíz cuadrada.

Si el valor de n es demasiado grande para calcular probabilidades utilizando la fórmula binomial o la tabla binomial (vuelve a los apartados anteriores de este capítulo), existe una alternativa. Si n es suficientemente grande, puedes utilizar la distribución normal para obtener una probabilidad binomial aproximada. La media y la desviación estándar intervienen en este proceso. Encontrarás los detalles en el capítulo 9.

Capítulo 9

La distribución normal

En este capítulo

- Comprender la distribución normal y la distribución normal estándar
- Ir de principio a fin al buscar probabilidades normales
- Ir hacia atrás para encontrar percentiles

En tus viajes estadísticos te encontrarás con dos grandes tipos de variables aleatorias: discretas y continuas. Las *variables aleatorias discretas* básicamente cuentan cosas (número de caras en diez lanzamientos de una moneda, número de mujeres demócratas en una muestra, etc.). La variable aleatoria discreta más conocida es la variable binomial (en el capítulo 8 encontrarás más información sobre las variables aleatorias discretas y las variables binomiales). Una *variable aleatoria continua* generalmente se basa en mediciones; puede abarcar un número de valores infinito no numerable (valores incluidos en un intervalo de la línea real) o puede tener tantos valores posibles como para considerarla continua (por ejemplo, el tiempo necesario para completar una tarea, las notas de un examen, etc.).

En este capítulo te explico y te enseño a calcular las probabilidades para la variable aleatoria continua más famosa de todos los tiempos: la distribución normal. También aprenderás a encontrar percentiles para la distribución normal (te dan una probabilidad en forma de porcentaje y tienes que averiguar el valor de *X* asociado a ella). Y piensa en lo graciosa que sería una camiseta donde pusiera "Los estadísticos son gente normal".

Fundamentos de la distribución normal

Una variable aleatoria continua *X* tiene una distribución normal si sus valores forman una curva continua acampanada. Cada distribución normal tiene su propia media, designada con la letra griega μ (se lee "mu"), y su propia desviación estándar, designada con la letra griega σ (se lee "sigma"). Pero con independencia de cuál sea la media y la desviación estándar, todas las distribuciones normales tienen la misma forma básica de campana. La figura 9-1 muestra algunos ejemplos de distribuciones normales.

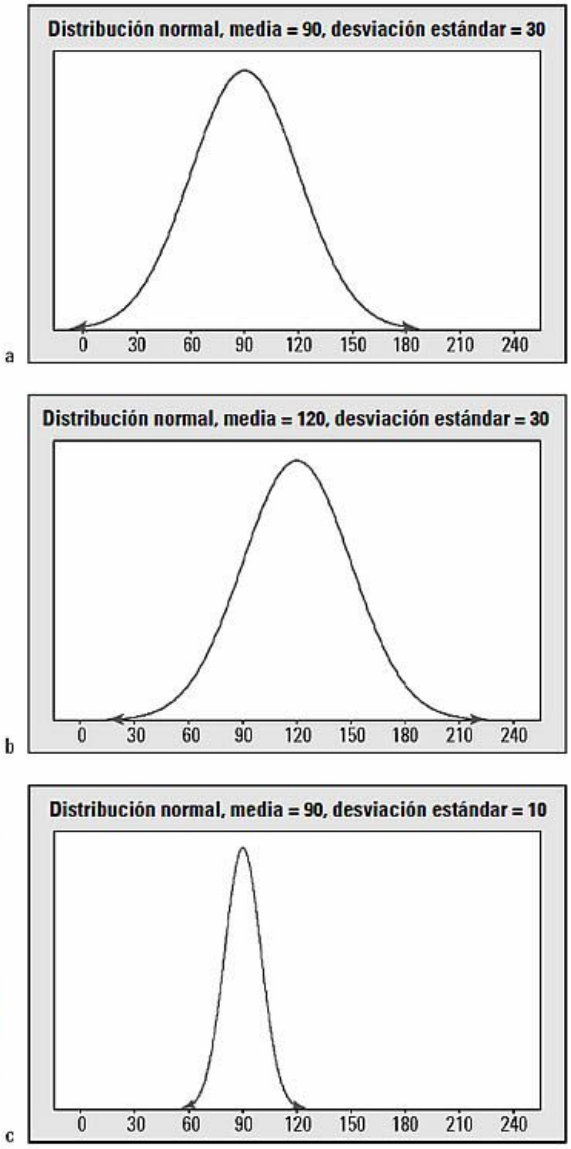


Figura 9-1: Tres distribuciones normales cuyas medias y desviaciones estándares son a) 90 y 30; b) 120 y 30; y c) 90 y 10, respectivamente

Todas las distribuciones normales tienen ciertas propiedades. Estas propiedades sirven para determinar la posición relativa de un resultado en concreto dentro de la distribución, y para encontrar probabilidades. Las propiedades de una distribución normal son las siguientes:

- ✔ Su forma es simétrica (es decir, si la cortas por el centro, la mitad izquierda es la imagen especular de la mitad derecha).
- ✔ Su distribución muestra una elevación en el centro, con colas que bajan por ambos lados.
- ✔ La media y la mediana son iguales y se encuentran exactamente en el centro de la distribución (debido a la simetría).
- ✔ Su desviación estándar es la distancia entre la media y el *punto de inflexión* (el lugar donde la curva pasa de ser convexa a cóncava).
- ✔ Debido a su característica forma acampanada, las probabilidades de la distribución normal cumplen la regla empírica (de la que encontrarás más información en el capítulo 5), que dice lo siguiente:
  - Aproximadamente el 68 % de sus valores se encuentran a no más de una desviación estándar respecto de la media. Para encontrar este rango, coge el valor de la desviación estándar y a continuación busca la media más esa cantidad y la media menos esa cantidad.
  - Aproximadamente el 95 % de sus valores se encuentran a no más de dos desviaciones estándares respecto de la media (en este caso coges dos veces la desviación estándar y luego sumas y restas esa cantidad desde la media).
  - Casi todos sus valores (aproximadamente el 99,7 % de ellos) se encuentran a no más de tres desviaciones estándares respecto de la media (coge tres veces la desviación estándar y luego sumas y restas esa cantidad desde la media).
- ✔ Es posible hallar probabilidades exactas para todos los intervalos de valores posibles en la distribución normal (no sólo para los que se encuentren a no más de 1, 2 o 3 desviaciones estándares respecto de la media) utilizando una tabla y sin realizar apenas ningún cálculo. (En el siguiente apartado encontrarás información detallada sobre esta tabla.)

Vuelve a mirar la figura 9-1. Si comparas las tres distribuciones mostradas en las figuras 9-1a, b y c, verás que todas ellas son simétricas y tienen la forma acampanada característica. Los ejemplos de la figura 9-1a y la figura 9-1b tienen la misma desviación estándar pero medias diferentes, y la figura 9-1b se encuentra 30 unidades a la derecha de la figura 9-1a (la diferencia entre una media de 120 y otra de 90). Las figuras 9-1a y c tienen la misma media (90), pero la figura 9-1a presenta más variabilidad que la figura 9-1c debido a que su desviación estándar es mayor (30, en comparación con 10). Debido a la mayor variabilidad, los valores de la figura 9-1a van desde 0 hasta 180 (aproximadamente), mientras que

los valores de la figura 9-1c sólo van desde 60 hasta 120.

Por último, las figuras 9-1b y c tienen diferentes medias y diferentes desviaciones estándares. La figura 9-1b tiene una media más alta que la desplaza hacia la derecha, y la figura 9-1c tiene una desviación estándar más pequeña, de manera que sus valores son los que están más concentrados en torno a la media.

Es importante fijarse en la media y en la desviación estándar para poder interpretar correctamente los números de una distribución normal. Por ejemplo, puedes comparar dónde se encuentra el número 120 en cada una de las distribuciones normales de la figura 9-1. En la figura 9-1a, el número 120 se encuentra una desviación estándar por encima de la media (como la desviación estándar es 30, resulta 90+1×30=120). Así pues, en esta primera distribución el número 120 es el valor más alto del rango donde se encuentran aproximadamente el 68% de los datos, según la regla empírica (puedes volver al capítulo 5).

En la figura 9-1b, el número 120 se encuentra exactamente en la media, el lugar donde los valores están más concentrados. En la figura 9-1c, el número 120 se encuentra casi en el extremo de la derecha, tres desviaciones estándares por encima de la media (como en este caso la desviación estándar es 10, resulta 90+3 [10]=120). En la figura 9-1c es muy improbable encontrar valores más altos que 120 porque estarían fuera del rango donde deben encontrarse aproximadamente el 99,7% de los valores, según la regla empírica.

La distribución normal estándar (distribución *Z*)

Un miembro muy especial de la familia de distribuciones normales es la distribución normal estándar, también llamada distribución *Z*. La *distribución Z* ayuda a encontrar probabilidades y percentiles para el resto de las distribuciones normales *(X)*. Sirve como referencia para el resto de distribuciones normales.

Particularidades de la distribución *Z*

La distribución *Z* es una distribución normal en la que la media es 0 y la desviación estándar es 1 (puedes ver un gráfico en la figura 9-2). Casi todos sus valores (el 99,7% aproximadamente) se encuentran entre –3 y +3, según la regla empírica. Los valores de la distribución *Z* se llaman valores *z*, puntuaciones *z* o puntuaciones estándares. Un valor *z* representa el número de desviaciones estándares que un determinado valor está por encima o por debajo de la media. Por ejemplo, *z*=1 en la distribución *Z* representa un valor que se encuentra a una desviación estándar por encima de la media. De forma similar, *z*=–1 representa un valor que se encuentra una desviación estándar por debajo de la media (lo indica el signo negativo del valor *z*). Un valor *z* igual a 0 se encuentra (lo has adivinado) justo en la media. Todos los valores *z* se reconocen universalmente.

Si vuelves a consultar la figura 9-1 y los comentarios sobre dónde se encuentra el número 120 en cada una de las distribuciones normales, dentro del apartado "Fundamentos de la distribución normal", ahora puedes calcular los valores *z* para tener una visión mucho más clara. En la figura 9-1a, el número 120 se encuentra una desviación estándar por encima de la media, de manera que su valor *z* es 1. En la figura 9-1b, el número 120 es igual a la media, de manera que su valor *z* es 0. En la figura 9-1c, el número 120 se encuentra tres desviaciones estándar por encima de la media, de manera que su valor *z* es 3.

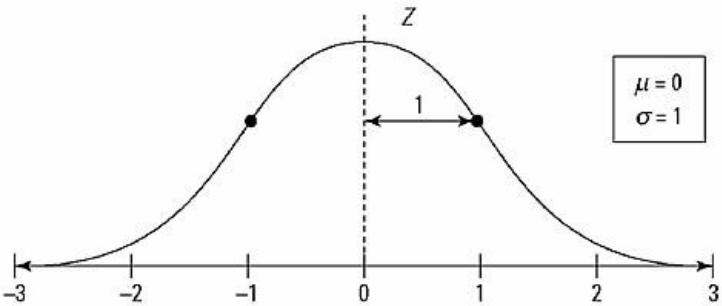


Figura 9-2: La distribución *Z* tiene media 0 y desviación estándar 1

No siempre es deseable que las puntuaciones estándar (valores *z*) sean altas. Por ejemplo, si estás midiendo el tiempo que tardas en dar una vuelta a la manzana corriendo, una puntuación estándar de +2 es mala porque significa que tu tiempo está dos desviaciones estándares por encima del tiempo medio (cuanto más alto es el tiempo, peor es la marca). En

este caso una puntuación estándar de –2 sería mucho mejor, pues indica un tiempo que está dos desviaciones estándares por debajo del tiempo medio (cuanto más bajo es el tiempo, mejor es la marca).

Normalizar de *X* a *Z*

Para hallar las probabilidades de una distribución continua tienes que determinar el área que hay bajo una curva (si dominas el cálculo, sabrás

que para eso hace falta integrar; si no dominas el cálculo, no te preocupes). Aunque trabajar con la campana de Gauss propia de una distribución normal parezca sencillo, el cálculo de áreas bajo esa curva requiere procedimientos matemáticos avanzados que pueden ser una auténtica pesadilla (y que escapan al alcance de este libro). Además, todas las distribuciones normales son diferentes, con lo que tienes que repetir este proceso cada vez que quieres encontrar una nueva probabilidad.

Para ayudarte a superar este obstáculo, los estadísticos hicieron todos los cálculos matemáticos para una distribución normal en concreto y elaboraron una tabla con todas sus probabilidades. ¿Adivinas qué distribución normal eligieron para hacer la tabla?

En efecto, todos los resultados básicos que necesitas para encontrar los valores de probabilidad de cualquier distribución normal *(X)* pueden reducirse a una sola tabla basada en la distribución normal estándar (distribución *Z*). Esta tabla se llama tabla *Z* y está incluida en el apéndice. Ahora ya sólo necesitas una fórmula que transforme los valores de una distribución normal *(X)* a la distribución *Z*; a partir de ahí podrás utilizar la tabla *Z* para encontrar cualquier probabilidad que quieras saber.

Convertir un valor *x* en un valor *z* se llama *normalizar*. La denominada "fórmula *z*" para convertir un valor *x* en un valor *z* es:

$$z = \frac{x - \mu}{\sigma}$$

Coges el valor *x*, le restas la media de *X* y luego divides por la desviación estándar de *X*. De este modo obtienes la puntuación estándar correspondiente (valor *z* o puntuación *z*).

Normalizar es como cambiar de unidades (por ejemplo, de grados Fahrenheit a grados centígrados). No afecta a las probabilidades de *X*. Por eso precisamente puedes utilizar la tabla *Z* para hallar esas probabilidades.

Con la fórmula *z* puedes normalizar un valor *x* de cualquier distribución (no sólo de la normal). De manera similar, no todas las puntuaciones estándares provienen de una distribución normal.

Como para normalizar restas la media al valor *x* y luego divides todo por la desviación estándar, literalmente estás dejando fuera de la ecuación la media y la desviación estándar de *X*. Esto es lo que te permite comparar todo lo que haya en la escala desde –3 hasta +3 (la distribución *Z*), donde los valores negativos indican una posición por debajo de la media, los valores positivos indican una posición por encima de la media, y el valor 0 indica que estás exactamente en la media.

Además, normalizar permite comparar números de distribuciones diferentes. Por ejemplo, imagina que Roberto saca un 80 (sobre 100) en el examen de matemáticas (que tiene una media de 70 y una desviación estándar de 10) y también en el examen de lengua (que tiene una media de 85 y una desviación estándar de 5). ¿Qué examen le salió mejor, en términos relativos, teniendo en cuenta el resto de la clase?

Si normalizas la nota obtenida por Roberto en matemáticas (80), te sale un valor *z* de $\frac{80-70}{10} = \frac{10}{10} = 1$. Por tanto, su nota en matemáticas está una desviación estándar por encima de la media de la clase. Si normalizas la nota que ha sacado en lengua (80), te sale un valor *z* de $\frac{80-85}{5} = \frac{-5}{5} = -1$, que sitúa a Roberto una desviación estándar por debajo de la media de la clase. Aunque Roberto haya sacado una nota de 80 en los dos exámenes, en realidad el examen de matemáticas le fue mejor que el de lengua (en términos relativos).

Para interpretar una puntuación estándar no necesitas conocer la puntuación original, la media ni la desviación estándar. La puntuación estándar te dice la posición relativa de un valor, que en la mayoría de los casos es lo más importante. De hecho, en la mayoría de las pruebas que se hacen para valorar el rendimiento académico de los alumnos, cuando te dan los resultados no te dicen siquiera cuál era la media y la desviación estándar; simplemente te dan tu puntuación *z* para que sepas en qué lugar de la distribución te encuentras.

Encontrar probabilidades para *Z* con la tabla *Z*

La tabla *Z* (tabla A-1 del apéndice) muestra probabilidades "menor que" para una amplia gama de valores *z*. Para encontrar probabilidades para la distribución normal estándar (distribución *Z*), haz lo siguiente:

1. **Ve a la fila que representa el dígito de la parte entera de tu valor *z* y el primer dígito tras la coma decimal.**

2. **Ve a la columna que representa el segundo dígito tras la coma decimal de tu valor *z*.**

3. **Busca la intersección entre fila y columna.**

Este resultado representa *p (Z <z)*, la probabilidad de que la variable aleatoria *Z* sea menor que el número *z* (o lo que es lo mismo: el porcentaje de valores *z* que son menores que el tuyo).

Por ejemplo, imagina que quieres encontrar *p* (*Z*<2,13). Vas a la tabla *Z* y buscas la fila correspondiente a 2,1 y la columna correspondiente a 0,03. En la intersección entre esa fila y esa columna encuentras la probabilidad 0,9834. Por tanto, *p* (*Z*<2,13)=0,9834.

Ahora pongamos que quieres encontrar *p* (*Z*<–2,13). Buscas la fila correspondiente a –2,1 y la columna correspondiente a 0,03. En la intersección de esa fila y esa columna se encuentra el valor 0,166. Por consiguiente, *p* (*Z*<–2,13) es 0,0166 (observa que esto es igual a uno menos la probabilidad de que *Z* sea menor que 2,13, ya que *p* (*Z*< +2,13) es 0,9834. Esto se cumple porque la distribución normal es simétrica. Encontrarás más información al respecto en el siguiente apartado).

Encontrar probabilidades para una distribución normal

Éstos son los pasos para encontrar una probabilidad cuando *X* tiene una distribución normal:

1. **Haz un dibujo de la distribución.**
2. **Traduce el problema a uno de los siguientes: *p (X<a)*, *p (X>b)* o *p (a<X<b)*. Sombrea el área correspondiente en el dibujo.**
3. **Normaliza *a* (o *b*) a una puntuación *z* utilizando la fórmula *z*:**

$$z = \frac{x - \mu}{\sigma}$$

4. **Busca la puntuación *z* en la tabla *Z* (la tabla A-1 del apéndice) y encuentra la probabilidad correspondiente.**

   (En el apartado "Normalizar de *X* a *Z*" encontrarás más información sobre la tabla *Z*.)

5a. Si necesitas una probabilidad "menor que", es decir, *p (X<a)*, ya has terminado.

5b. Si necesitas una probabilidad "mayor que", es decir, *p (X>b)*, resta a uno el resultado del paso 4.

5c. Si necesitas una probabilidad "entre dos valores", es decir, *p (a<X<b)*, sigue los pasos 1-4 para *b* (el mayor de los dos valores) y otra vez para *a* (el menor de los dos valores), y resta un resultado al otro.

La probabilidad de que *X* sea igual a un único valor es 0 para cualquier variable aleatoria continua (como la normal). La razón es que las variables aleatorias continuas consideran la probabilidad como el área que hay bajo la curva, y no hay ningún área bajo una curva en un único punto. Esto no pasa con las variables aleatorias discretas.

Imagina, por poner un ejemplo, que visitas un acuario y te quedas extasiado delante de un estanque de peces tropicales de colores, que tienen una distribución normal, con una media μ=16 centímetros y una desviación estándar σ=4 centímetros.

- ✔ Problema 1: ¿Cuál sería la probabilidad de atrapar un pez pequeño, digamos de menos de 8 centímetros?
- ✔ Problema 2: Supongamos que la organización quiere pasar los peces de más de 24 centímetros a otro estanque. ¿Cuál es la probabilidad de atrapar uno de estos especímenes?
- ✔ Problema 3: ¿Cuál sería la probabilidad de atrapar un pez de entre 16 y 24 centímetros?

Para resolver estos problemas siguiendo los pasos anteriores, empieza por dibujar la distribución normal que nos ocupa. La figura 9-3 muestra un gráfico de la distribución de *X* para las longitudes de los peces. Puedes ver dónde se encuentran los números de interés (8, 16 y 24).

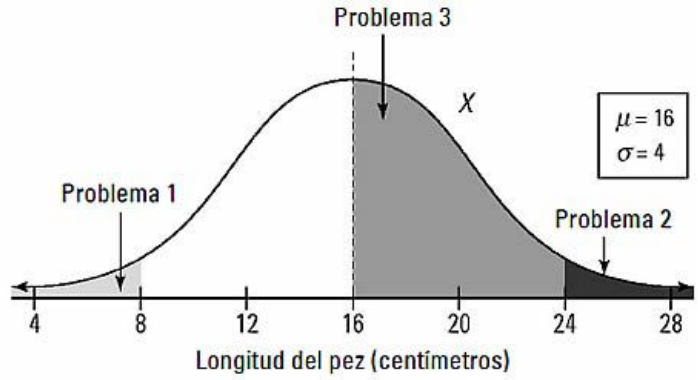


Figura 9-3: Distribución de las longitudes de los peces en un acuario

A continuación, traduce cada problema a la notación de probabilidad. El problema 1 en realidad te está pidiendo que encuentres p (X<8). Para el problema 2 quieres saber p (X>24). Y el problema 3 busca p (16<X<24).

El paso 3 te dice que conviertas los valores x en valores z utilizando la fórmula z:

$$z = \frac{x - \mu}{\sigma}$$

Para el problema 1 te queda lo siguiente:

$$p(X < 8) = p\left(Z < \frac{8-16}{4}\right) = p(Z < -2)$$

De forma similar, en el problema 2, $p(X>24)$ se convierte en

$$p(X > 24) = p\left(Z > \frac{24-16}{4}\right) = p(Z > 2)$$

Y en el problema 3, p (16<X<24) se convierte en

$$p(16 < X < 24) = p\left(\frac{16-16}{4} < Z < \frac{24-16}{4}\right) = p(0 < Z < 2)$$

La figura 9-4 muestra una comparación entre la distribución X y la distribución Z para los valores x=8, 16 y 24, que al normalizarse pasan a ser z=–2, 0 y +2, respectivamente.

Ahora que ya has convertido los valores x en valores z, continúas con el paso 4 y encuentras (o calculas) las probabilidades para esos valores z utilizando la tabla Z (incluida en el apéndice). En el problema 1 del ejemplo de los peces, quieres saber p (Z<–2), así que vas a la tabla Z, buscas la fila correspondiente a –2,0 y la columna correspondiente a 0,00, buscas la intersección entre ellas, y al final obtienes 0,0228. Según el paso 5a, ya has terminado. La probabilidad de pescar un pez de menos de 8 centímetros es igual a 0,0228.

El problema 2 te pide que encuentres p (Z> 2,00). Como buscas una probabilidad del tipo "mayor que", hay que seguir el paso 5b. Para poder utilizar la tabla Z, tienes que reformular esto con una expresión del tipo "menor que". Como la probabilidad total para la distribución Z es igual a 1, sabemos que p (Z>2,00)=1–p (Z<2,00)=1–0,9772=0,0228 (utilizando la tabla Z). Por tanto, la probabilidad de que un pez mida más de 24 centímetros también es 0,0228. (***Nota:*** las respuestas a los problemas 1 y 2 son iguales porque la distribución Z es simétrica; consulta la figura 9-3.)

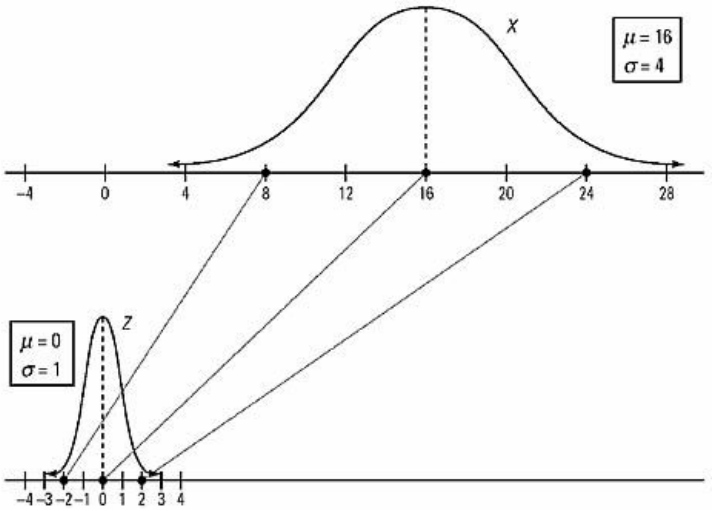


Figura 9-4: Normalización de números de una distribución normal (*X*) a números de la distribución *Z*

El problema 3 te pide que encuentres p (0<Z<2,00), y para eso hay que aplicar el paso 5c. Primero encuentras p (Z<2,00), que según la tabla Z es 0,9772. A continuación encuentras p (Z<0), que según la tabla Z es 0,5000. Si restas un valor del otro, 0,9772–0,5000=0,4772. La probabilidad de pescar un pez que mida entre 16 y 24 centímetros es 0,4772.

La tabla *Z* no contiene todos los valores posibles de *Z*, sino que llega tan sólo a los dos primeros dígitos tras la coma decimal. Utiliza el más próximo al que necesites. Igual que en un avión puede que la salida más próxima esté detrás de ti, puede que el valor *z* más próximo sea el que es inferior al que tú necesitas.

Encontrar *X* cuando conoces el porcentaje

Otro problema común sobre distribuciones normales consiste en encontrar percentiles para *X* (en el capítulo 5 hallarás una explicación detallada sobre los percentiles). Es decir, te dicen el porcentaje o probabilidad de estar en un determinado valor *x*, o por debajo de él, y tú tienes que encontrar el valor *x* correspondiente. Por ejemplo, si sabes que los golfistas cuyas tarjetas de puntuación estaban dentro del 10% con menos golpes consiguieron el pase al torneo, quizá te preguntes cuál era la puntuación de corte. Esa puntuación representa el percentil 10.

Un percentil no es un porcentaje. Un porcentaje es un número entre 0 y 100, mientras que un percentil es un valor de *X* (una altura, un coeficiente intelectual, una nota de examen, etc.).

Determinar un percentil para una distribución normal

Ciertos percentiles son tan populares que tienen sus propios nombres y su propia notación. Los tres percentiles "bautizados" son Q_1 (el primer cuartil, o percentil 25), Q_2 (el segundo cuartil, también llamado *mediana* o percentil 50) y Q_3 (el tercer cuartil o percentil 75). Encontrarás más información sobre los cuartiles en el capítulo 5.

Éstos son los pasos para encontrar cualquier percentil para una distribución normal *X*:

1a. Si la probabilidad (porcentaje) que te dicen es la probabilidad menor que *x* y tú tienes que encontrar *x*, lo traduces como: encuentra *a* si $p(X<a)=p$ (y *p* es la probabilidad que te dicen). Es decir, encuentra el percentil *p* para *X*. Ve al paso 2.

1b. Si la probabilidad (porcentaje) que te dicen es la probabilidad mayor que *x* y tú tienes que encontrar *x*, lo traduces como: encuentra *b* cuando $p(X>b)=p$ (y *p* es la probabilidad que te dicen). Reformúlalo como un problema de percentiles (del tipo "menor que"): encuentra *b* si $p(X<b)=1-p$. Es decir, encuentra el percentil $(1-p)$ para *X*.

2. Encuentra el percentil correspondiente para *Z*. Para ello, busca en el cuerpo de la tabla *Z* (incluida en el apéndice) la probabilidad que esté más cerca de *p* (para el paso 1a) o de $1-p$ (para el paso 1b). Busca la fila y la columna donde esté esa probabilidad (utilizando la tabla al revés). Ése es el valor *z* deseado.

3. Vuelve a convertir el valor *z* a un valor *x* (unidades originales) utilizando $x=\mu+z\sigma$. Ya tienes (¡por fin!) el percentil deseado para *X*.

La fórmula indicada en este paso no es más que una manera distinta de escribir la fórmula $z=\frac{x-\mu}{\sigma}$, despejando la *x*.

Un problema de percentil bajo

Volvamos al ejemplo de los peces utilizado anteriormente en el apartado "Encontrar probabilidades para una distribución normal", en la que las longitudes *(X)* de los peces de un acuario tienen una distribución normal con una media de 16 centímetros y una desviación estándar de 4 centímetros. Imagina que quieres saber qué longitud determina el 10% de peces más pequeños. ¿Qué percentil estás buscando?

Estar en el 10% más bajo significa que tienes una probabilidad "menor que" igual al 10%, y estás en el percentil 10.

Ahora ve al paso 1a del apartado anterior y traduce el problema. En este caso, como se trata de una situación del tipo "menor que", quieres encontrar *x* de forma que $p(X<x)=0{,}10$. Esto representa el percentil 10 para *X*.

La figura 9-5 representa gráficamente esta situación.

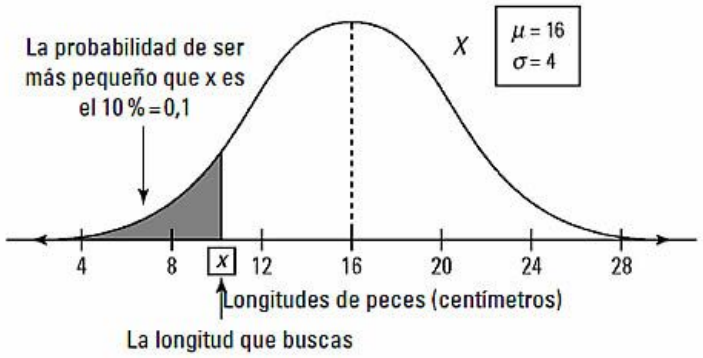


Figura 9-5: El 10 % de peces más pequeños del estanque

Ahora ve al paso 2, donde pone que debes encontrar el percentil 10 para *Z*. Si miras en el cuerpo de la tabla *Z* (en el apéndice), la probabilidad que está más cerca de 0,1 es 0,1003, situada en la intersección de la fila correspondiente a $z=-1,2$ y la columna correspondiente a 0,08. Por tanto, el percentil 10 para *Z* es –1,28; de manera que un pez cuya longitud esté a 1,28 desviaciones estándares por debajo de la media determinará el 10% de peces más pequeños (de entre todos los que hay en el estanque).

Pero ¿qué longitud tiene ese pez exactamente, en centímetros? En el paso 3 vuelves a convertir el valor *z* en un valor *x* (longitud en centímetros): utilizando la fórmula *z* con la *x* despejada, obtienes $x=16+-1,28\times4=10,88$ centímetros. Así pues, 10,88 centímetros es la longitud que determina el 10% de peces más pequeños. El 10 % de los peces del estanque miden menos que este valor.

Trabajar con un percentil alto

Ahora imagina que quieres saber qué longitud determina el 25% de peces más grandes en ese mismo estanque. Para resolver este problema tienes que seguir el paso 1b (del apartado "Determinar un percentil para una distribución normal"), ya que estar en la parte más alta de la distribución significa que estás buscando una probabilidad "mayor que". El número que buscas está en algún lugar de la cola derecha (la parte más alta) de la distribución *X*, con $p=25\%$ de la probabilidad a su derecha y $1-p=75$ % a su izquierda. Pensando en términos de la tabla *Z*, que únicamente utiliza probabilidades del tipo "menor que", tienes que encontrar el percentil 75 para *Z* y luego convertirlo a un valor *x*.

Paso 2: el percentil 75 de *Z* es el valor *z* para el que $p(Z<z)=0,75$. Utilizando la tabla *Z* (incluida en el apéndice), vemos que la probabilidad más próxima a 0,7500 es 0,7486, y el valor *z* correspondiente es la intersección de la fila para 0,6 y la columna para 0,07. Si juntas ambos, te sale un valor *z* de 0,67. Éste es el percentil 75 para *Z*. En el paso 3 vuelves a convertir el valor *z* en un valor *x* (longitud en centímetros): utilizando la fórmula *z* con la *x* despejada, obtienes $x=16+0,67\times4=18,68$ centímetros. Así pues, el 75% de los peces del estanque miden menos que 18,68 centímetros. Y para responder a la pregunta original, el 25% de peces más grandes del estanque miden más de 18,68 centímetros.

Problemas de percentiles difíciles de interpretar

Algunos problemas de percentiles son especialmente difíciles de interpretar. Por ejemplo, supongamos que el tiempo que tarda un caballo de carreras en dar una vuelta al hipódromo en una ronda de clasificación tiene una distribución normal con una media de 120 segundos y una desviación estándar de 5 segundos. Se clasifican el 10% de tiempos mejores, y el resto quedan fuera. ¿Cuál es el tiempo de corte para la clasificación?

Puesto que los mejores tiempos son los tiempos más bajos, el porcentaje de tiempos que están por debajo del corte debe ser 10, y el porcentaje de tiempos por encima del corte debe ser 90 (es fácil confundirse y pensar lo contrario). Por tanto, el percentil de interés es el 10, que es un valor bajo situado en la cola izquierda de la distribución. Ahora ya puedes resolver este problema igual que el problema 1 sobre longitudes de peces (está en el apartado "Encontrar probabilidades para una distribución normal"). En la tabla *Z* (incluida en el apéndice) vemos que la puntuación estándar para el percentil 10 es $z=-1,28$. Si convertimos de nuevo a unidades originales, tenemos que $x=\mu+z\sigma=120+(-1,28)(5)=113,6$ segundos. Por consiguiente, el tiempo de corte para que se clasifique un caballo de carreras (es decir, para estar entre el 10% de caballos más rápidos) es 113,6 segundos (observa que este número es inferior al tiempo medio de 120 segundos, lo cual tiene sentido; esto se debe a que el valor *z* es negativo).

El percentil 50 para la distribución normal es la media (debido a la simetría) y su puntuación *z* es cero. Los percentiles pequeños, como el 10, se encuentran por debajo de la media y tienen puntuaciones *z* negativas. Los percentiles grandes, como el 75, se encuentran por encima de la media y tienen puntuaciones *z* positivas.

Aquí tienes otro ejemplo de problema que podría despistarte un poco: imagina que el tiempo necesario para hacer un examen de estadística tiene una distribución normal, con una media de 40 minutos y una desviación estándar de 6 minutos. El tiempo de Vicente está en el percentil 90. ¿Qué porcentaje de alumnos continúan haciendo el examen cuando Vicente termina? Como Vicente está en el percentil 90, el 90% de los alumnos tardan menos que él en hacer el examen. Eso significa que el 90% de los alumnos entregan antes que Vicente, de manera que 100–90=10% de los alumnos continúan haciendo el examen cuando Vicente termina.

Para darte cuenta de que se trata de un problema de percentiles, busca pistas en el enunciado, por ejemplo que se mencione "el 10% más bajo" (también conocido como el percentil 10) o "el 10% más alto" (también conocido como el percentil 90). Si te hablan del 10% "mejor", tendrás que determinar si corresponde a las cifras más altas o a las más bajas.

Aproximación normal a la distribución binomial

Imagina que lanzas una moneda al aire 100 veces y *X* es el número de caras. ¿Cuál es la probabilidad de que *X* sea mayor que 60? En el capítulo 8 resolvimos problemas de este tipo (aunque con menos lanzamientos) utilizando la distribución binomial. Para los problemas binomiales en que *n* (el número de intentos) es pequeño, puedes utilizar la fórmula directa (la encontrarás en el capítulo 8) o la tabla binomial (la encontrarás en el apéndice), o bien recurrir a la tecnología, si está disponible (por ejemplo una calculadora gráfica o un programa de hoja de cálculo).

Sin embargo, cuando *n* es grande los cálculos se vuelven muy farragosos y la tabla binomial se queda sin números. Si no tienes a mano una herramienta tecnológica (como cuando estás haciendo un examen), ¿qué puedes hacer para encontrar una probabilidad binomial? Pues resulta que si *n* es suficientemente grande, puedes utilizar la distribución normal para obtener una respuesta muy aproximada con mucho menos trabajo.

Pero ¿qué quiero decir con que *n* sea "suficientemente grande"? Para determinar si *n* es suficientemente grande como para utilizar lo que los estadísticos denominan *aproximación normal a la probabilidad binomial*, deben cumplirse las dos condiciones siguientes:

- ✔ $n \times p \geq 10$ (mayor o igual que 10), donde *p* es la probabilidad de éxito.
- ✔ $n \times (1-p) \geq 10$ (mayor o igual que 10), donde $1-p$ es la probabilidad de fracaso.

Para encontrar la aproximación normal a la distribución binomial cuando *n* es grande, sigue los pasos siguientes:

1. **Asegúrate de que *n* es suficientemente grande como para utilizar la aproximación normal (comprueba si se cumplen las dos condiciones).**

   Para la pregunta sobre lanzamientos de una moneda, las condiciones se cumplen porque $n \times p = 100 \times 0{,}50 = 50$, y $n \times (1-p) = 100 \times (1-0{,}50) = 50$, y ambos valores son mayores o iguales que 10. Por tanto, puedes utilizar la aproximación normal.

2. **Expresa el problema en términos de la probabilidad de *X*.**

   En el ejemplo de la moneda, tienes que encontrar $p(X > 60)$.

3. **Normaliza el valor *x* a un valor *z* utilizando la fórmula *z*:**

   $$z = \frac{x - \mu}{\sigma}$$

   Para la media de la distribución normal, utiliza $\mu = np$ (la media de la distribución binomial), y para la desviación estándar σ, utiliza $\sqrt{np(1-p)}$ (la desviación estándar de la distribución binomial; puedes volver al capítulo 8).

   En el ejemplo de la moneda, utiliza $\mu = np = (100)(0{,}50) = 50$ y $\sigma = \sqrt{np(1-p)} = \sqrt{100(0{,}50)(1-0{,}50)} = 5$. Si ahora metes esos valores en la fórmula *z*, resulta $z = \frac{x - \mu}{\sigma} = \frac{60 - 50}{5} = 2$. Para resolver el problema tienes que encontrar $p(Z > 2)$.

Si en un examen te ponen un problema binomial, no te darán μ ni σ. Sin embargo, conoces las fórmulas que te permiten calcular ambos estadísticos utilizando *n* y *p* (ambas te las darán en el problema). Simplemente tienes que acordarte de calcular μ y σ (las necesitas para la fórmula *z*).

4. **Continúa igual que para cualquier distribución normal. Es decir, sigue los pasos 4 y 5 descritos en el apartado anterior "Encontrar probabilidades para una distribución normal".**

   Siguiendo con el ejemplo, en la tabla *Z* (incluida en el apéndice) vemos que $p(Z > 2{,}00) = 1 - 0{,}9772 = 0{,}0228$. Por consiguiente, la probabilidad de que salgan más de 60 caras al lanzar una moneda 100 veces es de tan sólo un 2,28 % aproximadamente (yo no apostaría a favor).

Cuando utilizas la aproximación normal para encontrar una probabilidad binomial, la respuesta es una aproximación (no es exacta). Asegúrate de

mencionarlo. Indica también que has comprobado las dos condiciones necesarias para utilizar la aproximación normal.

Capítulo 10

La distribución t

En este capítulo

- Características de la distribución *t*
- Relación entre la distribución *Z* y la distribución *t*
- Comprender y utilizar la tabla *t*

La distribución *t* es uno de los pilares del análisis de datos. A lo mejor has oído hablar de la prueba *t*, por ejemplo, que se utiliza a menudo para comparar dos grupos en estudios médicos y experimentos científicos. En este breve capítulo comento las principales características y aplicaciones de la distribución *t*. Explico la relación existente con la distribución normal (más información en el capítulo 9) y cómo utilizar la tabla *t* para encontrar probabilidades y percentiles.

Aspectos básicos de la distribución *t*

Este apartado ofrece una visión general de la distribución *t*, sus características principales, cuándo se utiliza y cuál es su relación con la distribución *Z* (ver el capítulo 9).

Comparación entre la distribución *t* y la distribución *Z*

La distribución normal es aquella que tiene la consabida forma de campana, con media μ y desviación estándar σ (encontrarás más información sobre la distribución normal en el capítulo 9). La distribución normal más frecuente es la normal estándar, también llamada distribución *Z*, en la que la media es 0 y la desviación estándar es 1.

La distribución *t* es similar a la distribución normal estándar. Ambas están centradas en el cero y tienen forma de campana, pero en comparación con la distribución *Z*, la *t* es más corta y más plana, y su desviación estándar es proporcionalmente mayor, razón por la cual las colas de ambos lados son más gruesas.

La figura 10-1 muestra una comparación entre la distribución *t* y la distribución *Z* en sus formas más genéricas.

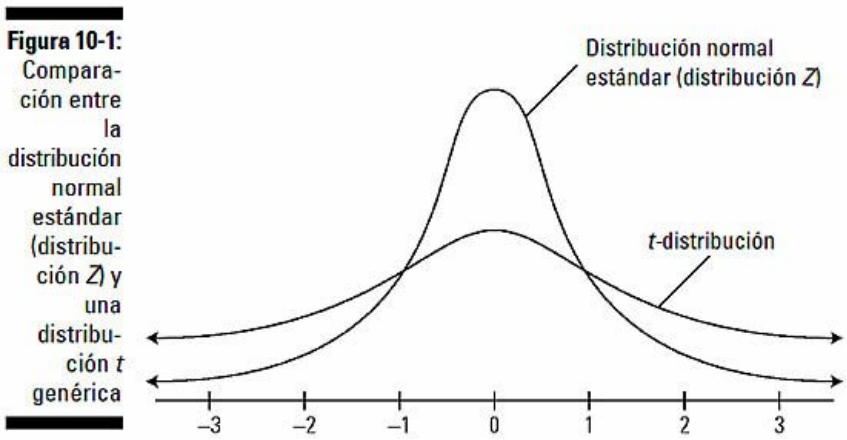


Figura 10-1: Comparación entre la distribución normal estándar (distribución *Z*) y una distribución *t* genérica

La distribución *t* generalmente se utiliza para estudiar la media de una población, no para estudiar los elementos de una población, En particular, se recurre a ella en muchos casos cuando se utilizan datos para estimar la media poblacional, por ejemplo para estimar el precio medio de todas las viviendas de nueva construcción en California. O también cuando se utilizan datos para verificar una afirmación referida a la media poblacional; por ejemplo, ¿es cierto que las viviendas de nueva construcción en California tienen un precio medio de 500.000 dólares?

Estos procedimientos se denominan *intervalos de confianza* y *contrastes de hipótesis*, y se tratan en los capítulos 13 y 14, respectivamente.

La relación entre la distribución normal y la distribución *t* es que la distribución *t* se utiliza a menudo para analizar la media de una población cuando dicha población tiene una distribución normal (o casi normal). Adquiere especial relevancia cuando el conjunto de datos es pequeño o cuando se desconoce la desviación estándar de la población (cosa que ocurre con frecuencia).

Cuando los estadísticos utilizan el término "distribución *t*", no se refieren a una única distribución. Existe una familia entera de distribuciones *t* específicas, en función del tamaño de muestra que se esté utilizando para estudiar la media poblacional. Cada distribución *t* se caracteriza por lo

que los estadísticos han dado en denominar "grados de libertad". Cuando tienes una población y el tamaño de muestra es *n*, los grados de libertad para la distribución *t* correspondiente son *n*–1. Por ejemplo, un tamaño muestral de 10 utiliza una distribución *t* con 10–1 (es decir, nueve) grados de libertad, lo cual se designa como t_9 (leído "te subnueve"). Cuando hay dos poblaciones se utilizan otros grados de libertad (se habla de ello en el capítulo 15).

Efecto de la variabilidad en las distribuciones *t*

Las distribuciones *t* basadas en tamaños muestrales pequeños tienen una desviación estándar mayor que las basadas en tamaños muestrales grandes. Su forma es más plana y sus valores están más dispersos. La razón es que los resultados basados en conjuntos de datos pequeños son más variables que los basados en conjuntos de datos grandes.

Cuanto más grande sea el tamaño muestral, mayor será el número de grados de libertad y más se parecerán las distribuciones *t* a la distribución normal estándar (distribución *Z*). Un punto de corte aproximado donde la distribución *t* y la distribución *Z* son similares es alrededor de *n*=30.

La figura 10-2 muestra el aspecto de varias distribuciones *t* con diferentes tamaños muestrales, comparados con la distribución normal estándar (distribución *Z*).

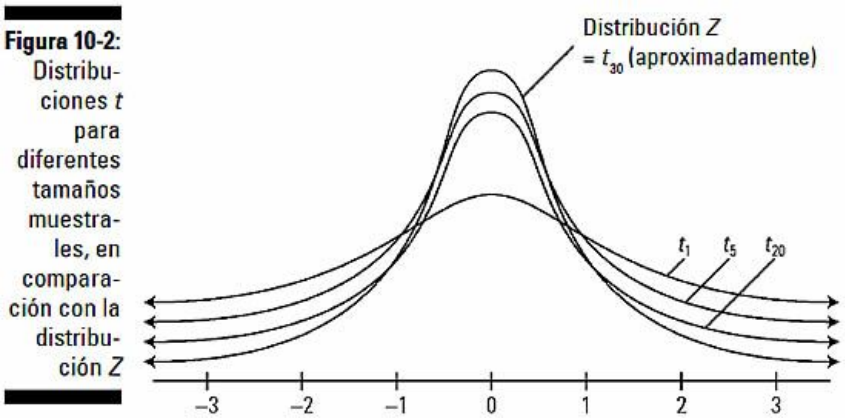


Figura 10-2: Distribuciones *t* para diferentes tamaños muestrales, en comparación con la distribución *Z*

Uso de la tabla *t*

Cada distribución normal tiene su propia media y desviación estándar que la caracterizan, de manera que buscar probabilidades para cada distribución normal no es un camino viable. Por suerte, puedes normalizar los valores de cualquier distribución normal y convertirlos en valores de una distribución normal estándar (distribución *Z*, cuya media es 0 y cuya desviación estándar es 1), y luego utilizar una tabla *Z* (incluida en el apéndice) para encontrar probabilidades. (En el capítulo hallarás más información sobre las distribuciones normales.)

Una distribución *t*, en cambio, no está caracterizada por su media y su desviación estándar, sino por el tamaño muestral del conjunto de datos utilizado *(n)*. Por desgracia, no existe una única distribución *t* estándar que puedas utilizar para transformar los números y encontrar probabilidades en una tabla. Como no sería humanamente posible crear una tabla de probabilidades y valores *t* correspondientes para todas las distribuciones *t* posibles, los estadísticos crearon una tabla que muestra ciertos valores de distribuciones *t* para algunos grados de libertad y algunas probabilidades. Esta tabla se llama tabla *t* y la hallarás en el apéndice. En este apartado te explico cómo encontrar probabilidades, percentiles y valores críticos (para intervalos de confianza) utilizando la tabla *t*.

Encontrar probabilidades con la tabla *t*

Cada fila de la tabla *t* (incluida en el apéndice) representa una distribución *t* diferente, caracterizada por sus grados de libertad *(gl)*. Las columnas representan diversas probabilidades comunes del tipo "mayor que", como por ejemplo 0,40, 0,25, 0,10 y 0,05. Los números dispuestos a lo largo de una fila son los valores de la distribución *t* (los valores *t*) correspondientes a las probabilidades "mayor que" indicadas en las cabeceras de las columnas. Las filas están ordenadas por grados de libertad.

Otro término para referirse a una probabilidad del tipo "mayor que" es *probabilidad de cola derecha*, lo cual indica que estas probabilidades representan áreas situadas en el extremo derecho (cola derecha) de la distribución *t*.

Por ejemplo, la segunda fila de la tabla *t* es para la distribución t_2 (dos grados de libertad, leído *te sub dos*). El segundo número (0,816) es el valor de la distribución t_2 cuya área a la derecha (su probabilidad de cola derecha) es 0,25 (ver la cabecera de la columna 2). Dicho de otro modo, la probabilidad de que t_2 sea mayor que 0,816 es 0,25. En notación probabilística esto es $p(t_2>0{,}816)=0{,}25$. El siguiente número de la fila dos de la

tabla *t* es 1,886, situado en la columna 0,10. Esto significa que la probabilidad de ser mayor que 1,886 en la distribución t_2 es 0,1. Como 1,886 está a la derecha de 0,816, su probabilidad de cola derecha es más baja.

Encontrar percentiles para la distribución *t*

La tabla *t* (incluida en el apéndice) también puede utilizarse para encontrar percentiles para una distribución *t*. Un *percentil* es un número de una distribución cuya probabilidad "menor que" es el porcentaje dado; por ejemplo, el percentil 95 de la distribución *t* con *n*–1 grados de libertad es el valor de t_{n-1} cuya probabilidad de cola izquierda (probabilidad "menor que") es 0,95 (y cuya probabilidad de cola derecha es 0,05). (Encontrarás más información sobre percentiles en el capítulo 5.)

Imagina que tienes una muestra de tamaño 10 y quieres encontrar el percentil 95 de la distribución *t* correspondiente. Tienes *n*–1 = 9 grados de libertad, así que buscas en la fila para *gl* = 9. El percentil 95 es el número por debajo del cual están el 95% de los valores y por encima del cual están el 5% de los valores, de manera que quieres que el área de la cola derecha sea 0,05. Avanzas por la fila hasta llegar a la columna correspondiente a 0,05 y encuentras que t_9=1,833. Éste es el percentil 95 de la distribución *t* que tiene 9 grados de libertad.

Al aumentar el tamaño muestral hasta *n*=20, el valor del percentil 95 disminuye; si miras en la fila correspondiente a 20–1=19 grados de libertad y en la columna 0,05 (una probabilidad de cola derecha igual a 0,05), verás que t_{19}=1,729. Observa que el percentil 95 para la distribución t_{19} es más pequeño que el percentil 95 para la distribución t_9 (1,833). La razón es que un número más alto de grados de libertad indica una desviación estándar más pequeña, de manera que los valores *t* están más concentrados en torno a la media y entonces alcanzas el percentil 95 con un valor de *t* más pequeño. (Ver el apartado "Efecto de la variabilidad en las distribuciones *t*" en este mismo capítulo.)

Escoger valores *t** para intervalos de confianza

Los *intervalos de confianza* estiman parámetros de una población, como la media poblacional, utilizando un estadístico (por ejemplo la media muestral) más/menos un margen de error. (En el capítulo 13 encontrarás toda la información que necesitas, y algo más, sobre los intervalos de confianza.) Para calcular el margen de error de un intervalo de confianza, necesitas un *valor crítico* (el número de errores estándares que sumas y restas para obtener el margen de error que quieres; ver el capítulo 13). Si el tamaño muestral es grande (por lo menos 30), utilizas valores críticos de la distribución *Z* (ver el capítulo 13) para construir el margen de error. Si el tamaño muestral es pequeño (menos que 30) o desconoces la desviación estándar de la población, utilizas la distribución *t* para encontrar valores críticos.

Para que te resulte más fácil encontrar valores para la distribución *t*, puedes utilizar la última fila de la tabla *t*, donde están los niveles de confianza más habituales, por ejemplo 80, 90 y 95%. Para encontrar un valor crítico, busca tu nivel de confianza en la fila de abajo del todo de la tabla; así sabrás qué columna de la tabla *t* necesitas. Busca la intersección entre esa columna y la fila correspondiente a tu número de grados de libertad. (En el capítulo 13 encontrarás fórmulas para los grados de libertad.) El número hallado es el valor crítico (o valor t^*) para tu intervalo de confianza. Por ejemplo, si quieres un valor t^* para un intervalo de confianza del 90% cuando tienes 9 grados de libertad, vas al final de la tabla, localizas la columna del 90% y buscas la intersección con la fila correspondiente a *gl*=9. El resultado es un valor t^* de 1,833 (redondeado).

La primera fila de la tabla *t* muestra probabilidades de cola derecha para la distribución *t*. Sin embargo, los intervalos de confianza requieren probabilidades de cola izquierda y probabilidades de cola derecha (porque sumas y restas el margen de error). Por tanto, a cada cola le corresponde la mitad de la probabilidad que queda después de restar el intervalo de confianza. Es importante que tengas esto en cuenta. Por ejemplo, un valor t^* para un intervalo de confianza del 90% tiene una probabilidad "mayor que" del 5% y una probabilidad "menor que" del 5% (coges el 100% menos el 90% y divides por 2). En la primera fila de la tabla *t* tendrías que buscar por 0,05 (en lugar de buscar el 10%, como quizá podrías verte inclinado a hacer), pero en la última fila de la tabla sencillamente buscas el 90% (el resultado obtenido con cualquiera de ambos métodos acaba estando en la misma columna).

Cuando busques valores t^* para intervalos de confianza, utiliza como guía la última fila de la tabla *t*, no las cabeceras de la parte de arriba de la tabla.

Estudiar comportamientos con la tabla *t*

Existen aplicaciones informáticas para calcular todas las probabilidades, percentiles o valores críticos que puedas necesitar para cualquier distribución *t* (o cualquier otra distribución), pero es probable que en los exámenes no puedas utilizarlas. De todos modos, una de las mayores ventajas de utilizar una tabla para encontrar probabilidades (en lugar de utilizar una aplicación informática) es que la tabla puede darte información sobre el comportamiento de la propia distribución, es decir, puede darte una perspectiva más amplia. A continuación comento algunas de las cosas que puedes averiguar sobre la distribución *t* analizando la tabla *t* (incluida en el apéndice).

En la figura 10-2 puedes ver que, a medida que los grados de libertad aumentan, los valores de cada distribución *t* se concentran más y más en torno a la media, hasta llegar a parecerse mucho a la distribución *Z*. La tabla *t* también confirma esta pauta. Debido a la forma en que está configurada la tabla *t*, si eliges una columna cualquiera y la recorres en sentido descendente, verás que los grados de libertad (y el tamaño muestral) aumentan y la probabilidad de cola derecha permanece igual. Simultáneamente los valores *t* se hacen cada vez más pequeños, lo cual

indica que están cada vez más cerca de la media (y, por tanto, más concentrados en torno a la media).

La penúltima fila de la tabla *t* está etiquetada con una *z* en la columna *gl*. Esto indica el "límite" de los valores *t* cuando el tamaño muestral *(n)* se hace infinito. Los valores *t* de esta fila son aproximadamente iguales que los valores *z* de la tabla *Z* (incluida en el apéndice) correspondientes a las mismas probabilidades "mayor que". Esto confirma lo que ya sabes: a medida que el tamaño muestral aumenta, la distribución *t* y la distribución *Z* son cada vez más parecidas. Por ejemplo, en la fila 30 de la tabla *t*, el valor *t* que corresponde a una probabilidad de cola derecha igual a 0,05 (columna 0,05) es 1,697. Este valor está muy cerca de $z=1,645$, el valor correspondiente a un área de cola derecha igual a 0,05 en la distribución *Z* (mira en la fila *Z* de la tabla *t*).

No hace falta un tamaño muestral enorme para que los valores de la distribución *t* se aproximen a los valores de una distribución *Z*. Por ejemplo, cuando $n=31$ y $gl=30$, los valores de la tabla *t* ya están muy cerca de los valores correspondientes de la tabla *Z*.

Capítulo 11

Distribuciones muestrales y el teorema del límite central

En este capítulo

- Comprender el concepto de una distribución muestral
- Ver la utilidad del teorema del límite central
- Determinar los factores que afectan a la precisión

Al tomar una muestra de datos, es importante darse cuenta de que los resultados varían de una muestra a otra. Los resultados estadísticos basados en muestras deberían incluir una medida de cuánto se espera que varíen. Cuando los medios de comunicación presentan estadísticas como el precio medio de un litro de gasolina o el porcentaje de viviendas vendidas durante el mes pasado, sabes que no incluyeron en su muestra todas las gasolineras del país ni todas las casas que se vendieron. La pregunta es la siguiente: ¿cuánto cambiarían esos resultados si se seleccionara una muestra distinta?

En este capítulo trataré de responder a esa pregunta estudiando el comportamiento de las medias y las proporciones de todas las muestras posibles. Estudiando el comportamiento de todas las muestras posibles puedes saber en qué lugar se encuentran los resultados de tu muestra y comprender qué significa que esos resultados se alejen de ciertas expectativas.

Definir una distribución muestral

Una *variable aleatoria* es una característica de interés que adopta ciertos valores de una manera aleatoria. Por ejemplo, el número de semáforos en rojo que te encuentras de camino a tu lugar de trabajo o centro de estudios es una variable aleatoria, lo mismo que el número de niños que hay en una familia elegida al azar. Utilizamos letras mayúsculas como *X* o *Y* para designar las variables aleatorias, y letras minúsculas como *x* o *y* para designar resultados concretos de variables aleatorias. Una *distribución* es una lista, gráfico o función de todos los resultados posibles de una variable aleatoria (por ejemplo *X*) y la frecuencia con la que ocurre cada resultado concreto *(x)* o conjunto de resultados. (En el capítulo 8 encontrarás más información sobre variables aleatorias y distribuciones.)

Por ejemplo, imagina que un millón de amigos tuyos lanzan un dado cada uno y anotan el resultado obtenido *(x)*. Una tabla o un gráfico de todos esos resultados posibles (del uno al seis) y la frecuencia con la que han ocurrido representa la distribución de la variable aleatoria *X*. La figura 11-1a muestra la distribución de *X* en este caso. Vemos que los números 1-6 aparecen todos con la misma frecuencia (cada uno ocurre 1 / 6 de las veces), que es lo que cabe esperar al cabo de muchos lanzamientos si el dado no está trucado.

Ahora supón que cada uno de tus amigos lanza su dado 50 veces (n=50) y anota el promedio, $\bar{x}$. El gráfico de todos los promedios de todas las muestras representa la distribución de la variable aleatoria $\overline{X}$. Como esta distribución está basada en medias muestrales, y no en resultados individuales, recibe un nombre especial. Se llama la *distribución muestral* de la media muestral, $\overline{X}$. La figura 11-1b muestra la distribución muestral de $\overline{X}$, el promedio de 50 lanzamientos de un dado.

La figura 11-1b (promedio de 50 lanzamientos) representa el mismo rango de resultados (del 1 al 6) que la figura 11-1a (lanzamientos individuales), pero la figura 11-1b tiene más resultados posibles. La razón es que puedes obtener un promedio de 3,3 o 2,8 o 3,9 puntos para 50 lanzamientos, por ejemplo, mientras que si lanzas un solo dado únicamente te puede salir un número entero del 1 al 6. Además, la forma de los gráficos es diferente: la figura 11-1a muestra una forma plana donde todos los resultados tienen la misma probabilidad, mientras que la figura 11-1b tiene forma de montículo, es decir, los resultados próximos al centro (3,5) ocurren con una frecuencia muy elevada y los resultados próximos a los extremos (1 y 6) ocurren con una frecuencia extremadamente baja. En los siguientes apartados analizaremos con detenimiento las diferencias y similitudes que presentan los elementos y los promedios en cuanto a forma, centro y dispersión, así como los motivos de tales diferencias y similitud. (Consulta antes el capítulo 8 si necesitas ponerte en antecedentes sobre la forma, el centro y la dispersión de variables aleatorias.)

La media de una distribución muestral

Retomando el ejemplo de los dados del apartado anterior, *X* es una variable aleatoria que indica el resultado que puedes obtener al lanzar un solo dado (suponiendo que ese dado no esté trucado). La media de *X* (para todos los resultados posibles) se designa como μ_X (se lee "mu subequis"), y en este caso su valor es 3,5 (como puedes ver en la figura 11-1a).

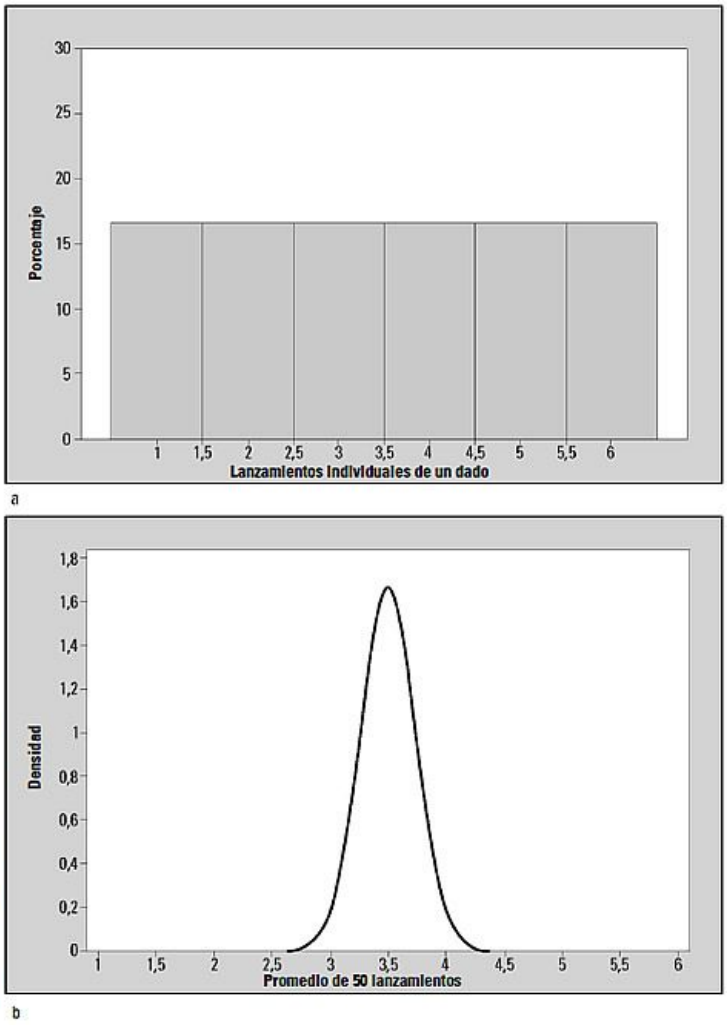


Figura 11-1: Distribuciones de a) lanzamientos individuales de un dado, y b) promedio de 50 lanzamientos de un dado

Si lanzas un dado 50 veces y calculas la media, la variable aleatoria $\overline{X}$ representa cualquier resultado que puedas obtener. La media de $\overline{X}$, designada como $\mu_{\overline{x}}$ (se lee "mu sub equis barra") también es 3,5 (puedes ver este resultado en la figura 11-1b).

¡Este resultado no es ninguna coincidencia! En general, la media de la población formada por todas las medias muestrales posibles es igual que la media de la población original (la notación matemática es $\mu_{\overline{x}} = \mu_x$). Es un poco enrevesado, pero tiene sentido que el promedio de los promedios de todas las muestras posibles coincida con el promedio de la población de donde vienen las muestras. En el ejemplo de los dados, el promedio de la población formada por todos los promedios de 50 lanzamientos es igual al promedio de la población formada por todos los lanzamientos individuales (3,5).

Utilizando subíndices para μ, puedes especificar de qué media estás hablando: la media de *X* (todos los elementos de una población) o la media de $\overline{X}$ (todas las medias muestrales de la población).

Medir el error estándar

Los valores de cualquier población se desvían de su media; por ejemplo, las alturas de las personas difieren de la altura media de toda la población. La variabilidad en una población de elementos *(X)* se mide en *desviaciones estándares*. (En el capítulo 5 encontrarás más información sobre la desviación estándar.) Las medias muestrales varían porque la muestra no abarca a toda la población, sino tan sólo un subconjunto; y como las muestras varían, las medias también. La variabilidad en la media muestral *($\overline{X}$)* se mide con *errores estándares*.

Aquí la palabra "error" no significa que alguien se haya equivocado. Significa que existe una divergencia entre la población y los resultados de la muestra.

El error estándar de la media muestral se designa como $\sigma_{\overline{x}}$ ("sigma subequis barra"). Su fórmula es $\frac{\sigma_x}{\sqrt{n}}$, en la que σ_x es la desviación estándar de la población ("sigma sub equis") y *n* es el tamaño de cada muestra. En los apartados siguientes veremos el efecto que tiene cada uno de estos dos componentes sobre el error estándar.

Tamaño muestral y error estándar

El primer componente del error estándar es el tamaño de la muestra, *n*. Como *n* se encuentra en el denominador de la fórmula para el error estándar, cuando *n* aumenta el error estándar disminuye. Si lo piensas, tiene sentido que al tener más datos exista menos variación (y más precisión) en los resultados. Pongamos que *X* es el tiempo que tarda un oficinista en escribir y enviar una carta de recomendación, y supongamos que *X* tiene una distribución normal con una media de 10,5 minutos y una desviación estándar de 3 minutos. La curva de abajo de la figura 11-2 representa la distribución de *X*, el tiempo individual de los oficinistas que componen la población. De acuerdo con la regla empírica (explicada en el capítulo 9), la mayoría de los valores se encuentran a no más de 3 desviaciones estándar respecto de la media (10,5), es decir, entre 1,5 y 19,5.

Ahora toma una muestra aleatoria de 10 oficinistas, mide sus tiempos y calcula la media, $\bar{x}$. Repite este proceso una y otra vez y representa gráficamente todos los resultados posibles para todas las muestras posibles. La curva central de la figura 11-2 representa la distribución muestral de $\overline{X}$. Observa que igualmente está centrada en 10,5 (cosa que ya esperabas) pero su variabilidad es menor; el error estándar en este caso es $\frac{\sigma_x}{\sqrt{n}} = \frac{3}{\sqrt{10}}$ = 0,95 minutos (bastante menos que 3 minutos, que era la desviación estándar del tiempo individual).

En la figura 11-2 puedes ver que el tiempo medio de las muestras de 10 oficinistas están más cerca de la media (10,5) que los tiempos individuales. La razón es que los tiempos medios no varían de una muestra a otra tanto como varían los tiempos individuales de una persona a otra.

Ahora toma todas las muestras aleatorias posibles de 50 oficinistas y calcula sus medias; la distribución muestral está representada por la curva de arriba de la figura 11-2. El error estándar de $\overline{X}$ se reduce hasta situarse en $\frac{\sigma_x}{\sqrt{n}} = \frac{3}{\sqrt{50}}$ = 0,42 minutos. Como puedes ver, el tiempo medio de 50 oficinistas está incluso más cerca de 10,5 que el tiempo medio de 10 oficinistas. Según la regla empírica, la mayoría de los valores están entre 10,5–3(0,42)=9,24 y 10,5+3(0,42)=11,76. Otras muestras más grandes proporcionan una precisión aún mayor en torno a la media porque cambian menos incluso de una muestra a otra.

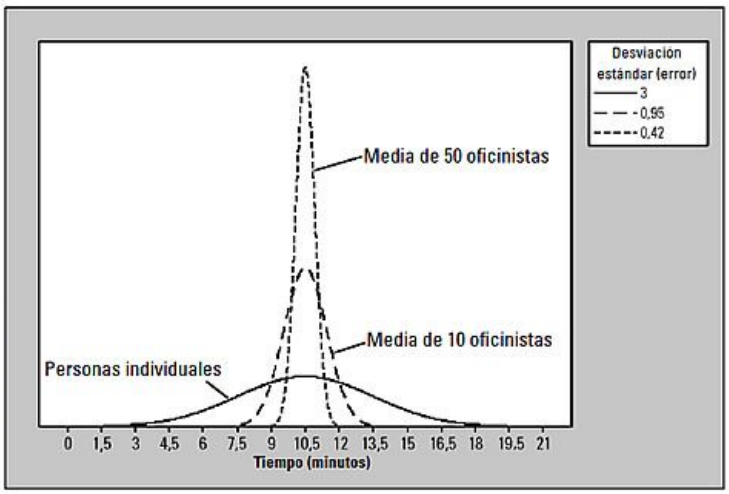


Figura 11-2: Distribuciones del tiempo de 1 oficinista, 10 oficinistas y 50 oficinistas

¿Por qué es importante tener una precisión mayor en torno a la media? Pues porque a veces no conoces la media pero quieres averiguar cuál es, o al menos acercarte lo más posible. ¿Cómo puedes hacerlo? Tomando una muestra aleatoria grande de la población y calculando su media. Si la muestra es muy grande, sabes que la media muestral será muy parecida a la media poblacional, tal y como puedes ver en la figura 11-2 (suponiendo que los datos se recopilen correctamente; en el capítulo 16 se explica la forma de recopilar datos válidos).

Desviación estándar de la población y error estándar

El segundo componente del error estándar se refiere a la diversidad existente en la población (medida mediante la desviación estándar). En la fórmula del error estándar $\frac{\sigma_x}{\sqrt{n}}$ para $\overline{X}$, se ve que la desviación estándar de la población, σ_x, se encuentra en el numerador. Esto significa que cuando aumenta la desviación estándar de la población, aumenta también el error estándar de las medias muestrales, lo cual tiene sentido desde el punto de vista matemático; ¿y desde el punto de vista de la estadística?

Imagina que tienes dos estanques llenos de peces (los llamaremos estanque 1 y estanque 2), y estás interesado en la longitud de los peces de cada estanque. Supongamos que las longitudes de los peces de cada estanque tienen una distribución normal (ver el capítulo 9). Te han dicho que las longitudes de los peces del estanque 1 tienen una media de 20 centímetros y una desviación estándar de 2 centímetros (figura 11-3a). Supongamos que los peces del estanque también tienen una media de 20 centímetros pero una desviación estándar más grande, de 5 centímetros (fíjate en la figura 11-3b).

Si comparas las figuras 11-3a y 11-3b, verás que las longitudes para ambas poblaciones de peces tienen la misma forma y la misma media, pero la distribución de la figura 11-3b (para el estanque 2) tiene más dispersión, o variabilidad, que la distribución representada en la figura 11-3a (para el estanque 1). Esta dispersión confirma que los peces del estanque 2 tienen longitudes más variables que los del estanque 1.

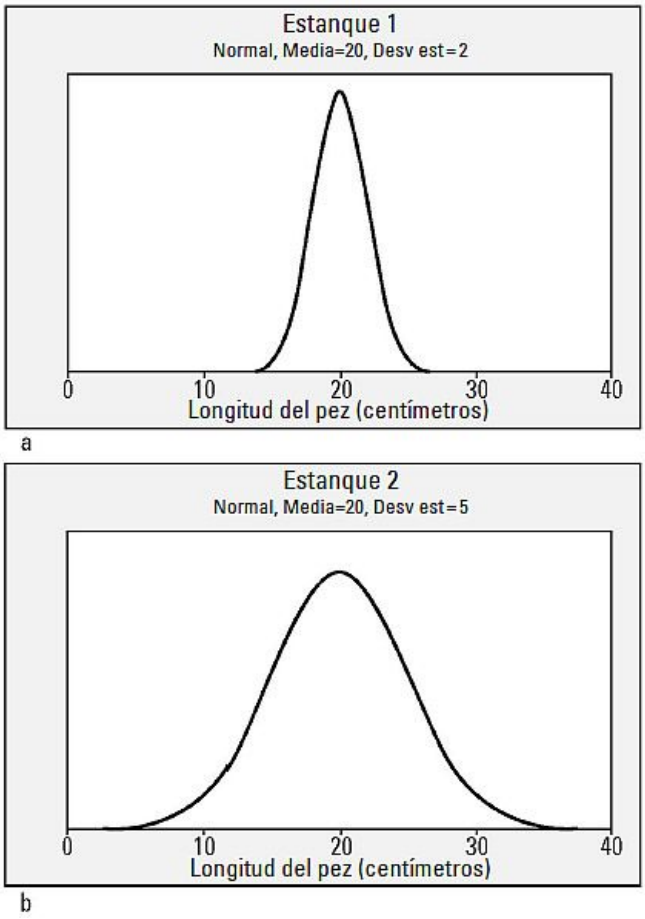


Figura 11-3: Distribuciones de las longitudes de los peces a) en el estanque 1; b) en el estanque 2

Ahora imagina que coges una muestra aleatoria de 100 peces del estanque 1, calculas la longitud media de los peces y repites este proceso una y otra vez. Y luego haces lo mismo con el estanque 2. Como las longitudes de los peces individuales del estanque 2 tienen más variabilidad que las longitudes de los peces individuales del estanque 1, sabes que las longitudes medias de las muestras del estanque 2 también tendrán más variabilidad que las longitudes medias de las muestras del estanque 1 (de hecho, aplicando la fórmula que te he dado anteriormente en este mismo apartado puedes calcular los errores estándares, que son 0,20 y 0,50, respectivamente).

Estimar la media poblacional es más difícil cuando la población varía mucho (resulta mucho más sencillo cuando los valores de la población son más uniformes). Con esto quiero decir que, cuando la desviación estándar de la población es mayor, el error estándar de la media muestral también es mayor.

Examinar la forma de una distribución muestral

Ahora que ya sabes encontrar la media y el error estándar de $\overline{X}$, el siguiente paso consiste en determinar la forma de la distribución muestral de $\overline{X}$; es decir, la forma de la distribución de todas las medias muestrales posibles (todos los valores posibles de $\bar{x}$) para todas las muestras posibles. Según las condiciones habrá que proceder de una manera u otra. Hay que distinguir entre dos casos: 1) la distribución original para *X* (la población) es normal, o tiene una distribución normal; y 2) la distribución original para *X* (la población) no es normal, o se desconoce.

Caso 1: la distribución de *X* es normal

Si *X* tiene una distribución normal, entonces $\overline{X}$ también, con independencia de cuál sea el tamaño muestral *n*. En el ejemplo sobre el tiempo *(X)* que tarda un oficinista en realizar una tarea (consulta el apartado "Tamaño muestral y error estándar"), sabías que *X* tenía una distribución normal (mira la curva de abajo del todo en la figura 11-2). Si te fijas en las otras curvas de la figura 11-2, verás que el tiempo medio para las muestras de $n=10$ y $n=50$ oficinistas, respectivamente, también tienen distribución normal.

Cuando *X* tiene una distribución normal, la medias muestrales también tienen siempre una distribución normal, sin importar qué tamaños muestrales cojas, incluso si eliges muestras de tan sólo 2 oficinistas cada vez.

La diferencia entre las curvas de la figura 11-2 no está en su media ni en sus formas, sino en el grado de variabilidad (lo cerca que están de la media los valores de la distribución). Los resultados basados en medias grandes varían menos y por eso están más concentrados en torno a la media que los resultados procedentes de muestras pequeñas o los resultados de los elementos que componen la población.

Caso 2: la distribución de *X* no es normal (entra en escena el teorema del límite central)

Si *X* tiene una distribución que no es normal, o si su distribución se desconoce, no puedes decir automáticamente que la media muestral *($\overline{X}$)* tiene una distribución normal. Sin embargo, y esto es increíble, puedes utilizar una distribución normal para hallar una aproximación a la distribución de $\overline{X}$, siempre y cuando el tamaño muestral sea suficientemente grande. Este resultado enormemente trascendente se debe a lo que los estadísticos han dado en llamar teorema del límite central.

El *teorema del límite central* (abreviado *TLC*) dice que si *X* no tiene una distribución normal (o su distribución se desconoce y, por tanto, no puede afirmarse que es normal), la forma de la distribución muestral de $\overline{X}$ es aproximadamente normal, suponiendo que el tamaño muestral, *n*, sea suficientemente grande. Es decir, se obtiene una aproximación a la distribución normal para las medias de muestras grandes, incluso si la distribución de los valores originales *(X)* no es normal.

La mayoría de los estadísticos están de acuerdo en que si *n* es al menos 30, esta aproximación estará razonablemente cerca de la distribución normal en la mayoría de los casos, aunque diferentes formas de distribución para *X* requieren diferentes valores de *n*. Cuanto mayor es el tamaño muestral (*n*), menos diferencia hay entre la distribución de las medias muestrales y una distribución normal.

La media de un dado no trucado sigue una distribución aproximadamente normal

Volvamos al ejemplo del dado, expuesto anteriormente en el apartado "Definir una distribución muestral". En la figura 11-1a puedes ver que la distribución de *X* (la población de resultados basados en millones de lanzamientos individuales) es plana; los resultados individuales de cada lanzamiento van de 1 a 6, y todos ellos son igual de probables.

Si consideramos los promedios, la cosa cambia. Si lanzas un dado un número de veces alto (pongamos una muestra de 50 veces) y miras los resultados, probablemente encuentres la misma cantidad de seises que de unos (observa que la media entre 6 y 1 es 3,5); de cincos que de doses (la media entre 5 y 2 también es 3,5); y de cuatros que de treses (la media también es 3,5; supongo que ya ves por dónde van los tiros). Por tanto, si lanzas un dado 50 veces, tienes una probabilidad elevada de que te salga una media próxima a 3,5. Algunas veces la casualidad hace que las cosas se tuerzan un poco, pero es raro que eso pase con 50 lanzamientos.

Obtener una media muy alta o muy baja con 50 lanzamientos es algo muy inusual. Para obtener una media de 1 con 50 lanzamientos, tiene que salirte un uno en todos los lanzamientos. ¿Qué probabilidad hay de que ocurra tal cosa? (si algún día te pasa, sal corriendo a comprar lotería, porque será el día más afortunado de toda tu vida). Y lo mismo se cumple para una media próxima a 6.

Por consiguiente, la probabilidad de que la media de 50 lanzamientos esté cerca del centro (3,5) es máxima, y la probabilidad de que esté cerca de los extremos (1 o 6) es muy pequeña. Igual que ocurre con las medias entre 1 y 6, las probabilidades disminuyen a medida que te alejas de 3,5, y aumentan a medida que te acercas a 3,5. En particular, como puedes ver en la figura 11-1b, la forma de la distribución muestral de medias muestrales es, aproximadamente, normal siempre y cuando el tamaño muestral sea suficientemente grande. (En el capítulo 9 encontrarás más información sobre la forma de la distribución normal.)

Ten presente que si lanzas el dado un número de veces incluso mayor, la probabilidad de que la media esté cerca de 3,5 aumenta, y la distribución muestral de las medias muestrales se parece cada vez más a una distribución normal.

La media de un dado trucado igualmente sigue una distribución aproximadamente normal

Sin embargo, a veces los valores de *X* no ocurren con la misma probabilidad que cuando lanzas un dado no trucado. ¿Qué ocurre entonces? Por ejemplo, supongamos que el dado está trucado y el valor medio de muchos lanzamientos individuales resulta ser 2, en lugar de 3,5. Esto significa que la distribución de *X* está sesgada a la derecha (más valores bajos como 1, 2 y 3, y menos valores altos como 4, 5 y 6). Pero si la distribución de *X* (millones de lanzamientos individuales de este dado trucado) está sesgada a la derecha; ¿cómo es posible que la distribución de $\overline{X}$ (promedio de 50 lanzamientos de este dado trucado) acabe siendo una distribución aproximadamente normal?

Pongamos que Roberto lanza el dado 50 veces. ¿Cómo será la distribución de los resultados? Es más probable que le salgan resultados bajos (como 1 y 2) y menos probable que le salgan resultados altos (como 5 y 6), de manera que la distribución de los resultados de Roberto también estará sesgada a la derecha.

De hecho, puesto que lanzó el dado muchas veces (50), la distribución de sus resultados individuales tiene una alta probabilidad de coincidir con la distribución de *X* (los resultados de millones de lanzamientos). Sin embargo, si tan sólo hubiera lanzado el dado unas pocas veces (pongamos 6 veces), es probable que números altos como el 5 y el 6 no le hubieran salido ni una sola vez, con lo que su distribución no se parecería en nada a la distribución de *X*.

Si miras los resultados de un millón de personas que, igual que Roberto, hayan lanzado este dado trucado 50 veces, todas esas distribuciones serán muy parecidas entre sí y muy parecidas a la distribución de *X*. Cuantos más lanzamientos haga este millón de personas, más cerca estarán sus distribuciones unas de otras y también de la distribución de *X*. Y aquí está la clave: si sus distribuciones de resultados tienen una forma similar, con independencia de cuál sea esa forma similar, entonces sus medias también serán similares. Algunas personas obtendrán medias más altas que 2 por casualidad, y otras obtendrán medias más bajas también por casualidad, pero esos tipos de medias se hacen cada vez menos probables cuanto más te alejas de 2. Esto significa que obtienes una distribución aproximadamente normal centrada en 2.

Lo importante de todo esto es que da igual si partes de una distribución sesgada o incluso de una distribución totalmente disparatada para *X*. Como los tamaños muestrales (número de lanzamientos) eran grandes, las distribuciones de los resultados de las distintas muestras acaban pareciéndose mucho, de manera que sus promedios serán similares, estarán cerca unos de otros y cerca de una distribución normal. En jerga estadística: la distribución de $\overline{X}$ es aproximadamente normal siempre y cuando *n* sea suficientemente grande. Y todo ello se deriva del teorema del límite central.

Para que el TLC funcione cuando *X* no tiene una distribución normal, cada persona tiene que lanzar el dado un número de veces suficiente (es decir, *n* debe ser suficientemente grande) como para que exista una buena probabilidad de que salgan todos los valores posibles de *X*, especialmente los resultados que ocurren con menos frecuencia. Si *n* es demasiado pequeño, habrá personas a quienes no les saldrán los resultados que tengan probabilidades más bajas, y entonces sus medias diferirán del resto más de lo que deberían. Como consecuencia, cuando juntes todas las medias, puede que no se concentren en torno a un único valor. En último término, puede que no aparezca la distribución aproximadamente normal.

Tres aclaraciones sobre el TLC

Para evitarte posibles confusiones, quiero puntualizar tres cosas sobre el teorema del límite central:

- ✔ El TLC únicamente es necesario cuando la distribución de*X* no es una distribución normal o bien se desconoce. Así pues, no es necesario cuando *X* comienza con una distribución normal.
- ✔ Las fórmulas para la media y el error estándar de $\overline{X}$ no son producto del TLC. Simplemente son resultados matemáticos que siempre se cumplen. Encontrarás estas fórmulas en los apartados "La media de una distribución muestral" y "Medir el error estándar", en este mismo capítulo.
- ✔ La *n* mencionada en el TLC se refiere al tamaño de la muestra que tomas cada vez, no al número de muestras que tomas. Si Roberto lanza el dado 50 veces, tienes una muestra de tamaño 50, de manera que *n*=50. Si diez personas hacen eso mismo, tienes 10 muestras, cada una de ellas con un tamaño de 50, de manera que *n* continúa siendo 50.

Encontrar probabilidades para la media muestral

Una vez que has confirmado, mediante las condiciones establecidas para el caso 1 o el caso 2 (si no los recuerdas puedes volver a leer los apartados anteriores), que $\overline{X}$ tiene una distribución normal o aproximadamente normal, estás de enhorabuena. La distribución normal es una distribución muy sencilla que tiene una tabla para encontrar probabilidades y cualquier otra cosa que desees. Por ejemplo, puedes encontrar probabilidades para $\overline{X}$ convirtiendo el valor $\bar{x}$ en un valor *z* y buscando probabilidades en la tabla *Z* (incluida en el apéndice). (En el capítulo 9 encontrarás más información sobre la distribución normal y la distribución *Z*.)

La fórmula general para convertir de valores $\bar{x}$ a valores *z* es:

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

Si sustituyes los valores apropiados de la media y el error estándar de $\overline{X}$,

la fórmula de conversión pasa a ser la siguiente:

$$z = \frac{\bar{x} - \mu_x}{\sigma_x / \sqrt{n}}$$

No te olvides de dividir por la raíz cuadrada de *n* en el denominador de *z*. Tienes que dividir siempre por la raíz cuadrada de *n* cuando la pregunta se refiere al *promedio* de los valores *x*.

Volviendo al ejemplo de los oficinistas del apartado anterior "Tamaño muestral y error estándar", imagina que *X* es el tiempo que un oficinista elegido al azar tarda en escribir y mandar una carta de recomendación estándar. Imagina que *X* tiene una distribución normal y que la media es 10,5 minutos y la desviación estándar es 3 minutos. Coges una muestra aleatoria de 50 oficinistas y mides el tiempo que tarda cada uno. ¿Cuál es la probabilidad de que su tiempo medio sea inferior a 9,5 minutos?

Para responder la pregunta tienes que encontrar P($\overline{X}$ <9,5). Para empezar, como *X* tiene una distribución normal, sabes que $\overline{X}$ también tiene una distribución exactamente normal (no aproximadamente normal). Si conviertes a *z*, resulta:

$$z = \frac{\bar{x} - \mu_x}{\sigma_x / \sqrt{n}} = \frac{9{,}5 - 10{,}5}{3 / \sqrt{50}} = -2{,}36$$

Por tanto quieres hallar *P(Z*<–2,36), que es igual a 0,0091 (según la tabla *Z* incluida en el apéndice). Así pues, la probabilidad de que una muestra aleatoria de 50 oficinistas tarde menos de 9,5 minutos, en promedio, en realizar esa tarea es del 0,91% (muy pequeña).

¿Cómo encuentras probabilidades para $\overline{X}$ si *X* no es normal o bien se desconoce? Según el TLC, la distribución de *X* puede ser no normal o incluso desconocida, y mientras *n* sea suficientemente grande podrás encontrar probabilidades aproximadas para $\overline{X}$ utilizando la distribución normal estándar (distribución *Z*) y el proceso anteriormente descrito. Es decir, convertir a un valor *z* y encontrar probabilidades aproximadas utilizando la tabla *Z* (incluida en el apéndice).

Cuando utilices el TLC para encontrar una probabilidad para $\overline{X}$ (es decir, cuando la distribución de *X* no sea normal o sea desconocida), asegúrate de mencionar que tu respuesta es una aproximación. También conviene mencionar que la respuesta aproximada es en realidad bastante exacta porque tienes una *n* suficientemente grande como para aplicar el TLC (si *n* no es suficientemente grande para el TLC, puedes utilizar la distribución *t* en muchos casos; está explicada en el capítulo 10).

Más allá de los cálculos, las probabilidades sobre $\overline{X}$ pueden ayudarte a decidir si una suposición o una afirmación referida a una media poblacional es correcta, según los datos que tienes. En el ejemplo de los oficinistas, partimos de que el tiempo medio que tardaban todos ellos en escribir una carta de recomendación era de 10,5 minutos. Sin embargo, tu muestra arrojaba una media de 9,5 minutos. Como viste que la probabilidad de que los oficinistas dieran un promedio inferior a 9,5 minutos era muy pequeña (0,0091), pudo ocurrir que al seleccionar tu muestra dieras por casualidad con un número inusualmente elevado de oficinistas muy diligentes, o puede que la suposición de que el tiempo medio de todos los oficinistas era 10,5 minutos sencillamente fuera demasiado alta (yo apuesto por esta segunda posibilidad). El proceso de verificar suposiciones o afirmaciones referidas a una población se denomina contraste de hipótesis. Encontrarás más información al respecto en el capítulo 14.

La distribución muestral de la proporción muestral

El teorema del límite central (TLC) no se aplica exclusivamente a medias muestrales para datos numéricos. También puede utilizarse con otros estadísticos, incluidas proporciones muestrales para datos categóricos (ver el capítulo 6). La *proporción poblacional*, *p*, es la proporción de elementos de la población que poseen una determinada característica de interés (por ejemplo, la proporción de ciudadanos que son votantes censados, o la proporción de adolescentes que tienen teléfono móvil). La *proporción muestral*, designada como $\hat{p}$ (se lee "pe sombrero"), es la proporción de elementos de la muestra que poseen esa característica en particular; o dicho de otro modo, el número de elementos de la muestra que tienen esa característica de interés dividido por el tamaño total de la muestra *(n)*.

Por ejemplo, si seleccionas una muestra de 100 adolescentes y compruebas que 60 de ellos tienen teléfono móvil, la proporción muestral de adolescentes con móvil es $\hat{p} = \frac{60}{100}$ = 0,6. Este apartado habla sobre la distribución muestral de todas las proporciones muestrales posibles, $\hat{p}$, a partir de muestras de tamaño *n* extraídas de una población.

La distribución muestral de $\hat{p}$ tiene las siguientes propiedades:

- ✔ Su media, designada como $\mu_{\hat{p}}$ (se lee "mu subpe sombrero"), es igual a la proporción poblacional, *p*.
- ✔ Su error estándar, designado como $\sigma_{\hat{p}}$ (se lee "sigma subpe sombrero"), es igual a:

$$\sqrt{\frac{p(1-p)}{n}}$$

  (Observa que, como *n* está en el denominador, cuando *n* aumenta, el error estándar disminuye.)

- ✔ Debido al TLC, su forma es aproximadamente normal, suponiendo que el tamaño de la muestra sea suficientemente grande. Por tanto, puedes utilizar la distribución normal para encontrar probabilidades aproximadas para $\hat{p}$.
- ✔ Cuanto mayor es el tamaño muestral *(n)*, menos diferencia hay entre la distribución de la proporción muestral y una distribución normal.

Cuando estás interesado en el número (y no en la proporción) de elementos de tu muestra que poseen una determinada característica, utilizas la distribución binomial para encontrar probabilidades para tus resultados (si lo necesitas vuelve al capítulo 8).

¿A partir de qué momento se puede considerar que el TLC funciona para proporciones muestrales? La mayoría de los estadísticos están de acuerdo en que tanto *np* como *n* (1–*p*) deben ser mayores o iguales que 10. Es decir, el número medio de éxitos *(np)* y el número medio de fracasos *n* (1–*p*) deben ser al menos 10.

Para entender mejor la distribución muestral de la proporción muestral, tomemos como ejemplo la encuesta que se entrega todos los años junto con las pruebas ACT (la selectividad estadounidense) para saber si el alumno quiere recibir ayuda en matemáticas. Supongamos (por investigaciones anteriores) que el 38% de todos los alumnos que hacen la prueba ACT responden que sí. Eso significa que *p*, la proporción poblacional, es igual a 0,38 en este caso. La distribución de respuestas (sí, no) para esta población aparece representada en la figura 11-4 en forma de gráfico de barras. (En el capítulo 6 encontrarás más información sobre los gráficos de barras).

Como el 38% se aplica a todos los alumnos que hacen el examen, para designar la proporción poblacional utilizo la notación *p* (en lugar de $\hat{p}$, que designa proporciones muestrales). Generalmente *p* es desconocida, pero aquí le asigno un valor para señalar cómo se comportan las proporciones muestrales en relación con la proporción poblacional.

Ahora tomemos todas las muestras posibles de *n*=1.000 alumnos de esta población y encontremos la proporción de alumnos de cada muestra que dijeron necesitar ayuda en matemáticas. La distribución de estas proporciones muestrales se representa en la figura 11-5. Tiene una distribución aproximadamente normal con una media *p*=0,38 y un error estándar igual a:

$$\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0{,}38(1-0{,}38)}{1.000}} = 0{,}015$$

(aproximadamente el 1,5 %).

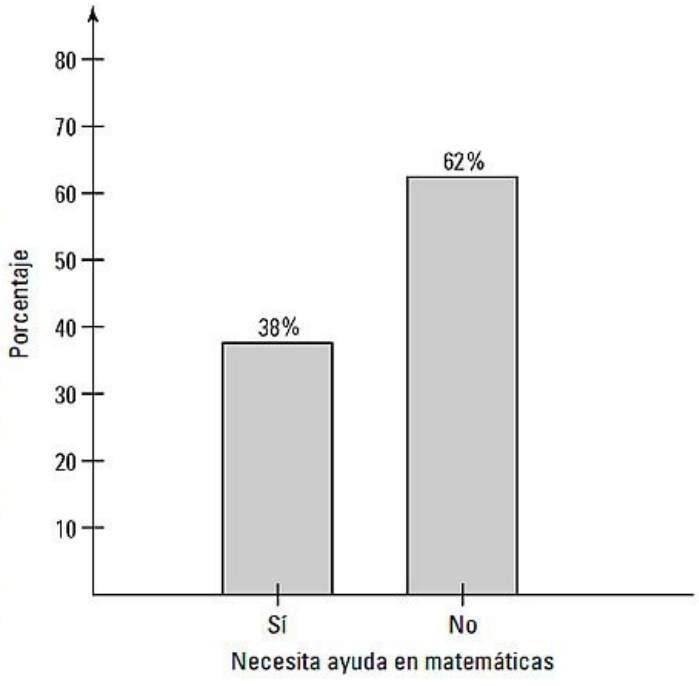


Figura 11-4: Porcentajes de población para las respuestas a la pregunta sobre ayuda en matemáticas (en la encuesta acompañante a las pruebas ACT)

La distribución aproximadamente normal funciona porque se cumplen las dos condiciones del TLC: 1) *np*=1.000 (0,38)=380 (≥ 10); y 2) *n* (1–*p*)=1.000 (0,62)=620 (también ≥ 10). Y como *n* es muy grande (1.000), la aproximación es excelente.

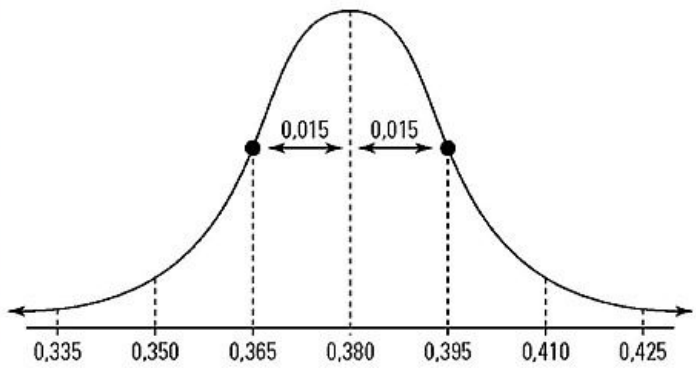


Figura 11-5: Distribución muestral de los alumnos que respondieron a la pregunta sobre ayuda en matemáticas, para muestras de tamaño 1.000

Encontrar probabilidades para la proporción muestral

Puedes encontrar probabilidades para $\hat{p}$, la proporción muestral, utilizando la aproximación normal siempre y cuando se cumplan las condiciones (descritas en el apartado anterior). En el ejemplo de las pruebas ACT, suponemos que el 38% de todos los alumnos que se examinan quieren recibir ayuda en matemáticas. Ahora imagina que coges una muestra aleatoria de 100 alumnos. ¿Cuál es la probabilidad de que más de 45 de esos alumnos digan que necesitan ayuda en matemáticas? Expresado como proporción del total, eso es equivalente a la probabilidad de que más de 45/100=0,45 de ellos digan que necesitan ayuda; es decir P ($\hat{p}$ > 0,45).

Para contestar esta pregunta, primero comprobamos las condiciones. En primer lugar, ¿*np* es mayor o igual que 10? Sí, porque 100×0,38=38. En segundo lugar, ¿*n* (1–*p*) es mayor o igual que 10? Otra vez sí, porque 100×(1–0,38)=62. Por tanto, puedes utilizar la aproximación normal.

Puedes convertir el valor $\hat{p}$ a un valor *z* utilizando la siguiente ecuación general:

$$z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Si utilizamos los números de este ejemplo, resulta:

$$z = \frac{0{,}45 - 0{,}38}{\sqrt{\frac{0{,}38(1-0{,}38)}{100}}} = 1{,}44$$

A continuación, utilizando la tabla A-1, que está en el apéndice, se ve que $P(Z > 1{,}44) = 1 - 0{,}9251 = 0{,}0749$. Por consiguiente, si es cierto que el 38% de los alumnos que hacen la prueba ACT necesitan ayuda en matemáticas, la probabilidad de coger una muestra aleatoria de 100 alumnos y encontrar más de 45 que necesiten ayuda en matemáticas es aproximadamente 0,0749 (en virtud del TLC).

Tal y como he mencionado en el apartado anterior sobre medias muestrales, puedes utilizar proporciones muestrales para verificar una afirmación sobre una proporción poblacional (este procedimiento es un contraste de hipótesis para una proporción poblacional; encontrarás información detallada en el capítulo 15). Volviendo al ejemplo de las pruebas ACT, la probabilidad de que más del 45% de los alumnos de una muestra de 100 necesiten ayuda en matemáticas (dando por sentado que el 38% de la población necesita ayuda en matemáticas) es 0,0749. Como esta probabilidad es mayor que 0,05 (el punto de corte típico para rechazar una afirmación referida a un valor poblacional), no puedes poner en duda que el porcentaje de la población que necesita ayuda en matemáticas es del 38%. Nuestro resultado muestral no llega a ser suficientemente poco común. (En el capítulo 15 encontrarás más información sobre contrastes de hipótesis para una proporción poblacional.)

Parte IV

Estimaciones aproximadas y formulación de hipótesis fiables

En esta parte...

Cuando te dan un estadístico aislado, no están contándote toda la película. Te falta saber la parte más importante: cuánto puede variar ese estadístico. Toda buena estimación de un parámetro poblacional incluye, además del estadístico, un margen de error. Esta combinación de un estadístico más/menos un margen de error se llama intervalo de confianza.

Ahora imagina que te dan una afirmación, suposición o valor objetivo para el parámetro poblacional y tú quieres verificar esa afirmación. Lo haces mediante un contraste de hipótesis basado en estadísticos muestrales. Como los estadísticos muestrales están sujetos a variación, necesitas técnicas que tengan eso en cuenta.

En esta parte encontrarás una explicación general e intuitiva del margen de error, los intervalos de confianza y los contrastes de hipótesis; su función, fórmulas, cálculos, factores influyentes e interpretación. Asimismo, podrás ver algunas referencias rápidas y ejemplos para los intervalos de confianza y los contrastes de hipótesis más utilizados.

Capítulo 12

Prever un margen de error

En este capítulo

- Comprender y calcular el margen de error
- Analizar el efecto del tamaño muestral
- Ver lo que no mide el margen de error

Los buenos investigadores siempre proporcionan alguna medida del grado de precisión de sus resultados, para que los consumidores de la información puedan poner esos resultados en perspectiva. Esta medida es lo que se denomina *margen de error (ME)*, una indicación de la divergencia que presumiblemente existe entre el estadístico muestral (un número que resume la muestra) y el parámetro poblacional objeto de estudio (un parámetro poblacional es un número que resume la población. Encontrarás más información sobre estadísticos y parámetros en el capítulo 4). Afortunadamente, muchos periodistas están empezando a darse cuenta de lo importante que es el ME para valorar la información, y por eso esta medida se menciona cada vez más en las noticias. Pero ¿qué es exactamente el margen de error? ¿Nos dice todo lo que necesitamos saber?

En este capítulo se examina el margen de error y lo que puede y no puede hacer para ayudarte a valorar la precisión de una información estadística. También se comenta la cuestión del tamaño muestral: te sorprendería lo pequeña que puede ser una muestra utilizada para tomar el pulso a la población de un país (o del mundo entero) si la investigación se lleva a cabo correctamente.

La importancia del más/menos

Imagino que el margen de error no es un término desconocido para ti. Es casi seguro que has oído hablar de él alguna vez, probablemente en el contexto de resultados de encuestas. Por ejemplo, quizá hayas oído a alguien decir: "Esta encuesta tiene un margen de error de más/menos 3 puntos porcentuales". Y puede que te has preguntado qué se supone que tienes que hacer con esa información y hasta qué punto es importante. La verdad es que los resultados de una encuesta por sí mismos (sin el ME) son tan sólo una medida de lo que pensaban las personas seleccionadas en una muestra; no reflejan cuál sería el sentir de la población entera si se hubiera preguntado a todos sus integrantes. El margen de error ayuda a estimar hasta qué punto los datos muestrales están cerca de la verdad sobre la población.

Los resultados basados en una muestra no serán exactamente iguales a los que habrías obtenido para la población entera, ya que al tomar una muestra no consigues información sobre todas las personas que forman la población. Sin embargo, si el estudio está bien hecho (en los capítulos 16 y 17 explico la forma de diseñar buenos estudios), los resultados de la muestra deberían ser representativos de los valores reales para toda la población, con un alto grado de confianza.

El ME no significa que alguien se haya equivocado; sólo indica que tu muestra no era toda la población y que, por ese motivo, los resultados difieren de esa población en cierta medida. Dicho de otro modo: admites que tus resultados cambiarán si se selecciona una muestra diferente y que únicamente son exactos dentro de un determinado intervalo de valores (que puede calcularse utilizando el margen de error).

Veamos, a modo de ejemplo, el tipo de encuesta utilizado por algunas de las principales empresas que realizan sondeos de opinión, como por ejemplo la organización Gallup. Supongamos que para el último sondeo seleccionaron una muestra de 1.000 ciudadanos estadounidenses, con el resultado de que 520 personas (el 52%) creen que su presidente lo está haciendo bien y 480 personas (el 48%) opinan lo contrario. Supongamos también que, según datos facilitados por la propia Organización Gallup, esta encuesta tiene un margen de error de más/menos el 3%. Sabes que la mayoría (más del 50%) de personas de esta muestra aprueban la gestión del presidente, pero ¿puedes afirmar que la mayoría de los estadounidenses también la aprueban? En este caso no puedes. ¿Por qué no?

Tienes que incluir el margen de error (en este caso, un 3%) en tus resultados. Si el 52% de las personas encuestadas (las seleccionadas para la muestra) aprueban la gestión del presidente, cabe esperar que el porcentaje de ciudadanos estadounidenses (la población entera) que opinarán eso mismo será el 52% más/menos el 3%. Por tanto, entre el 49 y el 55% de todos los estadounidenses aprueban la gestión del presidente. Eso es lo máximo que puedes acercarte con una muestra de 1.000 personas. Observa que el 49%, el extremo inferior de este intervalo, representa la minoría (porque es menos que el 50%). Por consiguiente, partiendo de esta muestra no puedes asegurar que la mayoría de los ciudadanos estadounidenses están a favor del presidente. Únicamente puedes decir que entre el 49 y el 55% de todos los estadounidenses están a favor del presidente, lo cual puede ser una mayoría y también puede no serlo.

Piensa un momento en el tamaño de la muestra. ¿No te parece interesante que una muestra de tan sólo 1.000 personas, para una población superior a 310 millones, te permita obtener unos resultados con un margen de error de tan solo más/menos el 3%? ¡A mí me parece increíble! Esto significa que, incluso con poblaciones muy grandes, basta con tomar una muestra minúscula para aproximarse al valor real (suponiendo,

como siempre, que los datos sean válidos). Como ves, la estadística es una herramienta poderosa para conocer la opinión de la gente sobre diversas cuestiones, y probablemente sea ésa la razón de que se hagan tantas encuestas y de que te importunen tantas veces para que las respondas.

Cuando trabajes con variables categóricas (las referidas a ciertas características que no requieren mediciones ni recuentos; ver el capítulo 6), una manera rápida de obtener una aproximación del margen de error para proporciones, para cualquier tamaño de muestra dado *(n)*, consiste simplemente en calcular 1 dividido por la raíz cuadrada de *n*. En el ejemplo de la encuesta Gallup, *n* es 1.000 y su raíz cuadrada es 31,62, de manera que el margen de error es aproximadamente 1 dividido entre 31,62, es decir, 0,03, que corresponde al 3%. En lo que queda de capítulo te explicaré cómo obtener una medición más precisa del margen de error.

Fórmula general para encontrar el margen de error

El margen de error es el valor "más/menos" que se añade al resultado de la muestra cuando dejas de referirte a dicha muestra y pasas a referirte a toda la población. Así pues, sabes que la fórmula general para el margen de error lleva delante el símbolo "±". Dicho esto, en este apartado te explico el procedimiento para determinar ese valor más/menos de manera precisa (no mediante el cálculo aproximado que hemos visto antes).

Medir la variabilidad de la muestra

Los resultados de la muestra varían, pero ¿en qué medida? Según el teorema del límite central (tratado en el capítulo 11), cuando el tamaño muestral es suficientemente grande, la denominada distribución muestral de las proporciones muestrales (o las medias muestrales) sigue una curva gaussiana (o una distribución aproximadamente normal, que puedes repasar en el capítulo 9). Algunas proporciones muestrales (o medias muestrales) sobrestiman el valor correspondiente a la población y otras lo subestiman, pero la mayoría están cerca del centro.

¿Y qué hay en el centro de esta distribución muestral? Si calculas el promedio de los resultados de todas las muestras posibles, ese promedio es la *proporción poblacional*, si se trata de datos categóricos, o la *media poblacional*, si se trata de datos numéricos. Normalmente no conoces todos los valores de la población, de manera que no puedes coger todos los resultados muestrales posibles y calcular la media, pero saber algo sobre el resto de las posibilidades muestrales sí te ayuda a medir cuánto esperas que varíe tu propia proporción (o media) muestral. (En el capítulo 11 encontrarás más información sobre medias y proporciones muestrales.)

Los errores estándares son los elementos básicos a partir de los cuales se obtiene el margen de error. El *error estándar* de un estadístico es básicamente la desviación estándar de la población dividida por la raíz cuadrada de *n* (el tamaño de la muestra). Esto refleja el hecho de que el tamaño de la muestra influye mucho en cuánto variará ese estadístico muestral de una muestra a otra. (En el capítulo 11 encontrarás más información sobre errores estándares.)

El número de errores estándares que tienes que sumar o restar para obtener el ME depende de la confianza que quieras tener en los resultados (lo que se denomina *nivel de confianza*). Por regla general se busca una confianza del 95%, de manera que la norma básica consiste en sumar o restar 2 errores estándares aproximadamente (1,96, siendo exactos) para obtener el ME (esto viene de la regla empírica, que tienes en el capítulo 9). De esta forma contemplas aproximadamente el 95% de todos los resultados posibles que pueden ocurrir si se toman muchas muestras. Para tener una confianza del 99% hay que sumar y restar 2,58 errores estándares (esto presupone una distribución normal con un valor *n* grande y una desviación estándar conocida; quizá necesites repasar el capítulo 11).

Puedes ser más preciso en cuanto al número de errores estándares que tienes que sumar o restar para calcular el ME con cualquier nivel de confianza; si las condiciones son las adecuadas, puedes utilizar valores de la distribución normal estándar (distribución *Z*). (Hay más información en el capítulo 13.) Para cualquier nivel de confianza dado, el valor correspondiente de la distribución normal estándar (llamado valor z^*) representa el número de errores estándares que tienes que sumar y restar para justificar ese nivel de confianza. Para un nivel de confianza del 95%, un valor z^* más exacto es 1,96 (es decir, "aproximadamente" 2), y para una confianza del 99%, el valor z^* exacto es 2,58. En la tabla 12-1 puedes ver algunos de los niveles de confianza (también llamados porcentajes de confianza) más utilizados, junto con sus valores z^* correspondientes.

Tabla 12-1. Valores z^* para varios niveles (porcentajes) de confianza escogidos

Porcentaje de confianza	*Valor z**
80	1,28
90	1,645
95	1,96
98	2,33
99	2,58

Para encontrar un valor z^* como los que aparecen en la tabla 12-1, incrementa el nivel de confianza para obtener la probabilidad "menor que" y luego encuentra su valor *z* correspondiente en la tabla *Z*. Por ejemplo, un nivel de confianza del 95% significa que la probabilidad "entre" es del 95%; por lo tanto, la probabilidad "menor que" es el 95% más el 2,5% (la mitad de lo que falta hasta el 100%), es decir, el 97,5%. Si buscas 0,975 en el cuerpo de la tabla *Z* , verás que $z^*=1,96$ para un nivel de confianza del 95%.

Calcular el margen de error para una proporción muestral

Cuando una pregunta de una encuesta pide a la gente que elija entre varias respuestas (por ejemplo: "¿Está usted a favor o en contra de la gestión del presidente?"), el estadístico utilizado para comunicar los resultados es la proporción de personas de la muestra que eligieron un determinado grupo (por ejemplo, el grupo "a favor"). Esto es lo que se llama *proporción muestral*. Para obtener esta cifra, coges el número de personas de la muestra que eligieron el grupo de interés y lo divides por el tamaño de la muestra, *n*.

Junto con la proporción muestral debes indicar el margen de error. La fórmula general del margen de error para la proporción muestral (si se cumplen ciertas condiciones) es $z^*\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, en la que $\hat{p}$ es la proporción muestral, *n* es el tamaño de la muestra, y z^* es el valor z^* correspondiente al nivel de confianza que buscas (según la tabla 12-1). Éstos son los pasos para calcular el margen de error para una proporción muestral:

1. **Encuentra el tamaño muestral, *n*, y la proporción muestral, $\hat{p}$**

   La proporción muestral es el número de la muestra que presenta la característica de interés, dividido por *n*.

2. **Multiplica la proporción muestral por (1 – $\hat{p}$).**

3. **Divide el resultado por *n*.**

4. **Calcula la raíz cuadrada de ese valor.**

   Ahora tienes el error estándar $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

5. **Multiplica el resultado por el valor z^* correspondiente al nivel de confianza deseado.**

   Este valor z^* correspondiente lo encontrarás en la tabla 12-1. Si el nivel de confianza es del 95%, el valor z^* es 1,96.

Volvamos al ejemplo de si los estadounidenses aprueban la gestión de su presidente y calculemos el margen de error. En primer lugar, supondremos que queremos un nivel de confianza del 95%, con lo que $z^*=1,96$. El número de estadounidenses de la muestra que dijeron estar de acuerdo con la gestión del presidente fue 520. Por tanto, la proporción muestral, , es 520/1.000=0,52 (el tamaño muestral, *n*, era 1.000.) El margen de error para esta pregunta se calcula de la manera siguiente:

$$z^*\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1,96\sqrt{\frac{(0,52)(0,48)}{1.000}}$$

$$= (1,96)(0,0158) = 0,0310$$

Según estos datos, concluimos, con un nivel de confianza del 95%, que el 52% de todos los estadounidenses (más/menos el 3,1%) están de acuerdo con la gestión de su presidente.

Para poder utilizar un valor z^* en la fórmula del margen de error para una proporción muestral deben cumplirse dos condiciones:

1. Debes estar seguro de que $n\hat{p}$ es mayor o igual que 10.

2. Debes estar seguro de que n (1 – $\hat{p}$) es mayor o igual que 10.

En el ejemplo anterior de la encuesta sobre la gestión presidencial, n = 1.000, $\hat{p}$ = 0,52, y 1 – $\hat{p}$ es 1 – 0,52 = 0,48. Ahora comprobemos las condiciones: $\hat{p}$ = 1.000 × 0,52 = 520, y n(1 - $n\hat{p}$) = 1.000 × 0,48 = 480. Los dos números son mayores o iguales que 10, así que no hay problema.

La mayoría de las encuestas se basan en cientos o incluso miles de personas, de manera que cumplir estas dos condiciones generalmente es pan comido (a menos que la proporción muestral sea muy grande o muy pequeña, en cuyo caso se requiere un tamaño muestral mayor para que las condiciones se cumplan).

Una proporción muestral es la versión decimal del porcentaje muestral. Dicho de otro modo, si tienes un porcentaje muestral del 5%, en la formula debes utilizar 0,05 (y no 5). Para convertir un porcentaje a su forma decimal, simplemente divide por 100. Una vez hayas hecho todos los cálculos, puedes volver a un valor porcentual multiplicando tu respuesta final por 100.

Comunicar resultados

Incluir el margen de error te permite extrapolar los resultados de la muestra a toda la población. Después de calcular e interpretar el margen de error, comunícalo junto con los resultados de tu encuesta. Si quisieras comunicar los resultados de la encuesta de opinión sobre la gestión presidencial, dirías lo siguiente: "Según los datos de mi muestra, el 52% de todos los estadounidenses aprueban la gestión de su presidente, con un margen de error de más/menos el 3,1%. Mi nivel de confianza en estos resultados es del 95%".

¿Cómo comunica sus resultados una empresa real que realiza sondeos de opinión? Aquí tienes un ejemplo de Gallup:

> *Según los datos de una muestra aleatoria de 1.000 adultos seleccionada para esta encuesta, podemos afirmar, con un nivel de confianza del 95%, que el margen de error de nuestro procedimiento de muestreo y de sus resultados no es superior a ±3,1 puntos porcentuales.*

Suena como esa larga lista de descargos de responsabilidad que aparece al final de un anuncio de *leasing* de vehículos. ¡Con la diferencia de que ahora sí entiendes la letra pequeña!

Nunca aceptes los resultados de una encuesta o estudio si no te dan el margen de error. El ME es la única forma de estimar la divergencia existente entre los estadísticos muestrales y los parámetros de población en los que estás interesado. Los resultados muestrales varían, de manera que si se hubiera seleccionado una muestra diferente, se habría obtenido un resultado muestral diferente; y el ME mide la magnitud de esta diferencia.

La próxima vez que en las noticias hablen de tal o cual encuesta o sondeo, fíjate a ver si te dicen el margen de error. Si no lo dicen, deberías preguntarte por qué. Los medios informativos ya suelen comunicar el margen de error para las encuestas, pero no ocurre lo mismo con otros tipos de estudios.

Calcular el margen de error para una media muestral

Cuando quieres estimar un parámetro a partir de una variable numérica (por ejemplo, la edad media de los profesores), el estadístico utilizado para ayudarte a estimar los resultados es el promedio de todas las respuestas facilitadas por las personas seleccionadas en la muestra. Esto es lo que se llama *media muestral* (o promedio muestral; está en el capítulo 5). Igual que para las proporciones muestrales, debes proporcionar el ME para las medias muestrales.

La fórmula general del margen de error para la media muestral (suponiendo que se cumple cierta condición) es $z^* \frac{\sigma}{\sqrt{n}}$, en la que σ es la desviación estándar de la población, n es el tamaño de la muestra, y z^* es el valor z^* correspondiente al nivel de confianza que buscas (según la tabla 12-1). Éstos son los pasos para calcular el margen de error para una media muestral:

1. **Encuentra la desviación estándar de la población, σ, y el tamaño muestral, *n*.**

   La desviación estándar de la población te la darán en el problema.

2. **Divide la desviación estándar de la población por la raíz cuadrada del tamaño muestral.**

   El error estándar es $\frac{\sigma}{\sqrt{n}}$.

3. Multiplica el resultado por el valor z^* correspondiente (consulta la tabla 12-1).

Por ejemplo, el valor z^* es 1,96 si quieres un nivel de confianza del 95%.

Para poder utilizar un valor z^* en la fórmula del margen de error para una media muestral debe cumplirse una de estas dos condiciones: 1) La población original tiene de entrada una distribución normal, o 2) El tamaño muestral es suficientemente grande como para poder utilizar la distribución normal (es decir, aplicamos el teorema del límite central; que explico en el capítulo 11). En general, el tamaño muestral, *n*, debe ser superior a 30 para el teorema del límite central. De todos modos, si es 29 tampoco te desesperes. El 30 no es un número mágico, sino más bien una regla general (en cualquier caso, debes conocer la desviación estándar de la población).

Imagina que eres el gerente de una heladería y estás enseñando a un grupo de nuevos empleados a llenar los cucuruchos grandes con la cantidad correcta de *topping* de virutas de chocolate (10 gramos). Quieres estimar el peso medio del *topping* de chocolate que utilizan durante un día, incluido el margen de error. En lugar de pesar todas las virutas, pides a cada empleado que compruebe los pesos de una muestra aleatoria de *topping* y anote dichos pesos en un cuaderno. Para una muestra de $n=50$ helados, la media muestral resulta ser 10,3 gramos. Supongamos que la desviación estándar de la población, σ, es 0,6 gramos.

¿Cuál es el margen de error? (suponiendo que quieres un nivel de confianza del 95%). Se calcula de la siguiente forma:

$$z^* \frac{\sigma}{\sqrt{n}} = 1{,}96 \frac{0{,}6}{\sqrt{50}} = (1{,}96)(0{,}0849) = 0{,}17$$

Para comunicar estos resultados, dirás que, según los datos de la muestra de 50 helados, estimas que el peso medio del *topping* de chocolate que han puesto los nuevos empleados durante un día es 10,3 gramos, con un margen de error de ±0,17 gramos. En otras palabras, se estima (con un nivel de confianza del 95%) que el intervalo de valores probables para el peso medio de todos los *topping* llenados durante ese día está comprendido entre 10,30–0,17=10,13 gramos y 10,30+0,17=10,47 gramos. Al parecer, los nuevos empleados están demasiado generosos con las virutas de chocolate (aunque supongo que los clientes no se van a quejar).

Observa que en el ejemplo de los cucuruchos de helado las unidades son gramos, no porcentajes. Cuando trabajes con datos, y también cuando comuniques resultados, recuerda siempre cuáles son las unidades. Además, asegúrate de que las estadísticas se comunican con las unidades correctas, y en caso contrario pregunta cuáles son esas unidades.

En los casos en que *n* es demasiado pequeña (en general, inferior a 30) para poder utilizar el teorema del límite central, pero tú igualmente crees que los datos proceden de una distribución normal, puedes utilizar un valor t^* en lugar de un valor z^* en las fórmulas. Un valor t^* es el que proviene de una distribución *t* con $n-1$ grados de libertad (en el capítulo 10 encontrarás información detallada sobre la distribución *t*). De hecho, muchos estadísticos utilizan habitualmente valores t^* en lugar de valores z^*, puesto que, si el tamaño muestral es grande, los valores t^* y los valores z^* son aproximadamente iguales. Asimismo, si la desviación estándar de la población, σ, es desconocida, puedes sustituirla por *s*, la desviación estándar de la muestra (en las fórmulas también utilizas un valor t^* en lugar de un valor z^*).

Cuestión de confianza

Si quieres un nivel de confianza en los resultados superior al 95%, tendrás que sumar y restar más de 1,96 errores estándares (observa la tabla 12-1). Por ejemplo, para tener una confianza del 99%, sumas y restas 2,58 errores estándares para obtener el margen de error. Sin embargo, más confianza significa un margen de error más grande (suponiendo que el tamaño muestral permanezca invariable); así que debes preguntarte si merece la pena. Si pasas de un nivel de confianza del 95% a uno del 99%, el valor z^* aumenta 2,58–1,96=0,62 (mira la tabla 12-1). La mayoría de la gente opina que no merece la pena sumar y restar esta cantidad de error para aumentar sólo el 4% (99% frente a 95%) el nivel de confianza en los resultados obtenidos.

Nunca puedes estar completamente seguro de que los resultados muestrales reflejen la realidad de la población, ni siquiera si incluyes el margen de error. Aunque tengas una confianza del 95% en tus resultados, eso significa que si repites el proceso de muestreo una y otra vez, el 5% de las veces la muestra no representará bien a la población, simplemente debido al azar (no debido a problemas con el proceso de muestreo ni por cualquier otro motivo). En esos casos tus resultados estarían equivocados. Ten esto siempre presente.

Repercusiones del tamaño muestral

Las dos ideas más importantes en relación con el tamaño muestral y el margen de error son las siguientes:

> ✔ El tamaño muestral y el margen de error tienen una relación inversa.
>
> ✔ A partir de cierto punto, incrementar *n* deja de ser tan provechoso.

A lo largo de este apartado explicaré estos dos conceptos.

Tamaño muestral y margen de error

La relación entre el margen de error y el tamaño muestral es simple: cuando el tamaño muestral aumenta, el margen de error disminuye. Esta relación se califica de inversa porque ambos van en direcciones opuestas. Si lo piensas, tiene sentido que cuanta más información tengas, más exactos serán los resultados (o, dicho de otro modo, más pequeño será el margen de error; suponiendo, claro está, que los datos se recopilen y manipulen correctamente).

En el apartado anterior hemos visto que la consecuencia de un nivel de confianza más elevado es que el ME será mayor. Sin embargo, aumentando el tamaño muestral es posible compensar ese ME mayor y reducirlo hasta un tamaño razonable. Encontrarás más información al respecto en el capítulo 13.

No por ser más grande es (mucho) mejor

En el ejemplo del sondeo sobre el índice de aprobación del presidente (planteado en el apartado anterior "Calcular el margen de error para una proporción muestral"), los resultados de una muestra de tan sólo 1.000 personas reflejan, con una desviación del 3%, aproximadamente, lo que hubiera contestado la población entera (310.000.000 de ciudadanos estadounidenses) si les hubieran preguntado a todos ellos.

Utilizando la fórmula del margen de error para una proporción muestral, puedes ver cómo el margen de error cambia drásticamente para muestras de diferentes tamaños. Volviendo al sondeo sobre el índice de aprobación del presidente, supongamos que *n* fuera 500 en lugar de 1.000 (recuerda que $\hat{p}$ =0,52 para este ejemplo). Entonces el margen de error para una confianza del 95% es $1{,}96\sqrt{\frac{(0{,}52)(0{,}48)}{500}} = (1{,}96)(0{,}0223) = 0{,}0438$, es decir, el 4,38%.

Cuando *n*=1.000 en el mismo ejemplo, el margen de error (para una confianza del 95%) es $1{,}96\sqrt{\frac{(0{,}52)(0{,}48)}{1000}} = (1{,}96)(0{,}0158) = 0{,}0310$, que corresponde al 3,10%. Si se aumenta *n* hasta 1.500, el margen de error (con el mismo nivel de confianza) pasa a ser $1{,}96\sqrt{\frac{(0{,}52)(0{,}48)}{1500}} = (1{,}96)(0{,}0129) = 0{,}0253$, o el 2,53%. Finalmente, cuando *n* = 2.000, el margen de error es $1{,}96\sqrt{\frac{(0{,}52)(0{,}48)}{2000}} = (1{,}96)(0{,}0112) = 0{,}0219$, o el 2,19%.

A la luz de todos estos resultados puedes ver que un tamaño muestral mayor se reduce el ME, pero a partir de cierto punto la mejora no es tanta. Por cada persona más que incluyes en la encuesta, el coste aumenta, y pasar el tamaño muestral de, pongamos, 1.500 a 2.000 disminuye el margen de error tan sólo el 0,34% (¡una tercera parte del 1%!), de 0,0253 a 0,0219. En muchos casos el sobrecoste y los inconvenientes asociados a esa pequeña reducción del ME no merecen la pena. De ahí que digamos que no por ser más grande es mucho mejor.

Pero lo que quizá te sorprenda es que, a la postre, por ser más grande puede ser incluso peor. En el apartado siguiente te explico esta sorprendente revelación.

No perder la perspectiva del margen de error

El margen de error indica hasta qué punto los resultados de la muestra son representativos de la población entera que es objeto de estudio (o al menos proporciona un límite superior del error que cabe esperar). Como las conclusiones sobre la población están basadas en una única muestra, es preciso indicar cuánto pueden variar esos resultados muestrales como resultado del mero azar.

Otra visión del margen de error es que representa la distancia máxima esperada entre los resultados muestrales y los resultados de toda la población (si hubieras podido obtenerlos por medio de un censo). Por supuesto, si conocieras la verdad absoluta sobre la población no intentarías hacer una encuesta, ¿verdad?

Tan importante como saber lo que el margen de error mide es comprender lo que el margen de error no mide. El margen de error únicamente mide la variación aleatoria. Es decir, no mide ningún tipo de sesgo ni errores que puedan ocurrir durante la selección de los participantes, la preparación o realización de la encuesta, el proceso de recopilación e introducción de datos, el análisis de esos datos y la extracción de las conclusiones finales.

Un buen lema que puedes recordar cuando analices resultados estadísticos es el siguiente: “Si entra basura, sale basura”. Por muy bonito y científico que sea el margen de error, recuerda que la fórmula utilizada para calcularlo no dice nada sobre la calidad de los datos en los que está basado. Si la proporción muestral o la media muestral estaba basada en una *muestra sesgada* (una que favorezca a determinadas personas en detrimento de otras), un diseño equivocado, un procedimiento de recopilación de datos erróneo, preguntas sesgadas o errores sistemáticos en el registro de los resultados, entonces calcular el margen de error no sirve de nada porque carece de significado.

Por ejemplo, una cifra de 50.000 encuestados suena estupendamente, pero si resulta que todas ellas eran visitantes de un determinado sitio web, el margen de error para ese resultado será erróneo porque el cálculo estará basado en resultados sesgados. De hecho, muchas muestras extremadamente grandes son el resultado de procedimientos de muestreo sesgados. Por supuesto, hay quien sigue adelante y hace públicos los resultados a pesar de todo, de manera que te corresponde a ti descubrir qué utilizaron para la fórmula: información válida o basura. Si resulta ser basura, ya sabes lo que debes hacer con el margen de error: no hacerle caso. (Para más información sobre los errores que pueden ocurrir durante una encuesta o experimento, consulta los capítulos 16 y 17, respectivamente.)

Cuando presenta los resultados de una encuesta, Gallup incluye una declaración de descargo donde se aborda la cuestión de lo que el margen de error mide y no mide. Gallup dice que, además del error de muestreo, las encuestas pueden tener otros errores o estar sesgadas debido a la manera de formular las preguntas y a determinados problemas logísticos asociados a la realización de encuestas (por ejemplo, que falten datos de números de teléfono dados de baja).

Esto significa que, incluso con la mejor de las intenciones y la máxima atención al detalle y al control del proceso, hay cosas que salen mal. No hay nada perfecto. Lo importante es saber que el margen de error no puede medir el alcance de esos otros tipos de errores. Y si una empresa con tanta credibilidad como Gallup admite la posibilidad de sesgo, ya puedes imaginar lo que ocurre realmente con otros estudios que ni de lejos están tan bien diseñados y realizados.

Capítulo 13

Intervalos de confianza: estimaciones aproximadas

En este capítulo

- Comprender los elementos y la interpretación de un intervalo de confianza
- Realizar cálculos con confianza
- Analizar los factores que influyen en la amplitud de un intervalo de confianza
- Detectar resultados engañosos

La mayoría de los estadísticos se utilizan para estimar alguna característica sobre una población de interés, por ejemplo los ingresos medios por unidad familiar, el porcentaje de personas que compran regalos de cumpleaños por Internet, o la cantidad media de helado que consume un español en un año (y el aumento de peso resultante, por desgracia). Estas características de una población se llaman *parámetros*. Por regla general, para estimar el valor de un parámetro se toma una muestra de la población y se utilizan estadísticos de esa muestra que permitan obtener una estimación válida. La pregunta es la siguiente: ¿qué es una estimación válida?

Si el proceso se lleva a cabo correctamente (cosa que no suelen hacer los medios de comunicación), la estimación estará muy cerca del parámetro. Este capítulo ofrece una visión general de los intervalos de confianza (el tipo de estimaciones utilizado y recomendado por los estadísticos): por qué deben utilizarse (en lugar de una estimación consistente en un único número), cómo definir, calcular e interpretar los intervalos de confianza más utilizados, y cómo detectar estimaciones engañosas.

No todas las estimaciones son iguales

Basta con leer una revista o un periódico o escuchar un programa informativo para oír montones de estadísticas, muchas de las cuales son estimaciones de alguna cantidad. Quizá te preguntes de dónde salen esas estadísticas. En algunos casos las cifras son el resultado de una investigación bien realizada, y en otros casos no son más que un disparo a ciegas. Aquí tienes, a modo de ejemplo, varias estimaciones que encontré en un solo número de una revista sobre negocios publicada en Estados Unidos. Proceden de diversas fuentes:

- ✔ Aunque en determinados sectores cuesta colocarse, hay otros sectores donde se busca gente: en los próximos ocho años se necesitarán 13.000 auxiliares de anestesia. Los sueldos oscilan entre 80.000 y 95.000 dólares al año.
- ✔ Un jugador de la liga profesional de béisbol utiliza una media de 90 bates cada temporada.
- ✔ El Lamborghini Murciélago puede pasar de 0 a 100 kilómetros por hora en 3,7 segundos y alcanza una velocidad máxima de casi 330 kilómetros por hora.

Algunas de estas estimaciones son más fáciles de obtener que otras. Pongo a continuación algunas observaciones referidas a esas estimaciones:

- ✔ ¿Cuál es la forma de estimar cuántos auxiliares de anestesia harán falta en los próximos ocho años? Podrías empezar por averiguar cuántos se jubilarán en ese plazo, pero eso no tendría en cuenta el aumento de la demanda. Podríamos predecir de manera más o menos fiable la oferta de empleos para el próximo año o dos años, pero a ocho años vista es mucho más difícil.
- ✔ El número medio de bates utilizados por un jugador de la liga profesional a lo largo de una temporada podría hallarse preguntando a los propios jugadores, a las personas que se encargan del equipamiento o a las empresas que suministran los bates.
- ✔ Determinar la aceleración de un coche es más difícil, pero podría hacerse en una pista con ayuda de un cronómetro. Habría que medir la aceleración media de muchos coches (no sólo uno) de la misma marca y modelo, siempre bajo las mismas condiciones de conducción.

No todas las estadísticas son iguales. No te creas de entrada todo lo que te digan. Para determinar si una estadística es fiable y creíble, piensa si tiene sentido y qué harías tú para obtener una estimación. Si la estadística es muy importante para ti, averigua qué procedimiento siguieron para obtenerla. (En el capítulo 16 encontrarás todo lo relativo a las encuestas, y en el capítulo 17 te doy pelos y señales sobre los experimentos.)

Vincular un estadístico a un parámetro

Un *parámetro* es un único número que describe a una población, por ejemplo los ingresos medios de una unidad familiar en Estados Unidos. Un

estadístico es un único número que describe a una muestra, por ejemplo los ingresos medios de una unidad familiar en una muestra de, pongamos, 1.200 unidades familiares. Lo normal es que desconozcas los valores de parámetros poblacionales, de manera que tomas muestras y utilizas estadísticos para dar la estimación posible.

Supongamos que quieres saber qué porcentaje de todos los vehículos de Estados Unidos son camionetas *pickup* (ése es el parámetro, en este caso). No puedes comprobar todos los vehículos, así que tomas una muestra aleatoria de 1.000 vehículos en varias autopistas a diferentes horas del día. Observas que el 7% de los vehículos de tu muestra son camionetas *pickup*. Llegados a este punto, no puedes afirmar que exactamente el 7% de todos los vehículos que circulan por las carreteras estadounidenses son camionetas *pickup*, porque sabes que esto se basa exclusivamente en los 1.000 vehículos que componen tu muestra. Aunque confías en que ese 7% esté cerca del porcentaje real, no puedes estar seguro porque tus resultados se basan en una muestra de vehículos, no en todos los vehículos de Estados Unidos.

Entonces, ¿qué haces? Pues coges el resultado de tu muestra y le sumas y restas un número para indicar que estás dando un intervalo de valores posibles para el parámetro poblacional, en lugar de dar por sentado que el estadístico muestral es igual al parámetro poblacional (lo cual no sería correcto, aunque los medios de comunicación lo hacen continuamente). Ese número que sumas y restas al estadístico se llama *margen de error (ME)*. El más/menos (se escribe ±) que se añade a la estimación ayuda a ver los resultados con perspectiva. Si conoces el margen de error, te haces una idea de cuánto podrían variar los resultados muestrales si tomaras otra muestra.

La palabra "error" (en el término "margen de error") no significa que alguien se haya equivocado ni que los datos sean inválidos. Simplemente quiere decir que los resultados de una muestra no son exactamente iguales a los que habrías obtenido si hubieras incluido a toda la población. Esta divergencia mide el error debido a la probabilidad aleatoria, el puro azar, no el debido al sesgo (por eso es tan importante reducir el sesgo al mínimo al seleccionar la muestra y al recopilar los datos; ver los capítulos 16 y 17).

Terminología básica

Un estadístico más/menos un margen de error es lo que se llama un *intervalo de confianza*:

- ✔ La palabra *intervalo* se utiliza porque el resultado consiste en un intervalo de valores. Por ejemplo, supongamos que al 40% de los niños les gusta el fútbol, con un margen de error de más/menos el 3,5%. Eso significa que el porcentaje de niños a los que les gusta el fútbol está entre el 40%–3,5%=36,5% y el 40%+3,5%=43,5%. El extremo inferior del intervalo es el estadístico menos el margen de error, y el extremo superior es el estadístico más el margen de error.
- ✔ Con todos los intervalos de confianza tienes cierto grado de confianza en que has estimado correctamente el parámetro por medio de tu muestra. Este grado de confianza se expresa como porcentaje y se denomina *nivel de confianza*.

Más adelante en este mismo capítulo encontrarás fórmulas y ejemplos para los intervalos de confianza más utilizados.

A continuación verás los pasos generales para estimar un parámetro con un intervalo de confianza. En lo que queda de capítulo te proporciono información pormenorizada sobre los pasos 1 y 4-6. Los pasos 2 y 3 se refieren al muestreo y la recopilación de datos, que se explican con más detalle en el capítulo 16 (muestreo y recopilación de datos para encuestas) y en el capítulo 17 (recopilación de datos de experimentos):

1. **Elegir el nivel de confianza y el tamaño muestral.**
2. **Seleccionar una muestra aleatoria de elementos de la población.**
3. **Recopilar datos fiables y relevantes de los elementos que componen la muestra.**
4. **Resumir los datos en un estadístico, por ejemplo una media o una proporción.**
5. **Calcular el margen de error.**
6. **Sumar y restar el margen de error al estadístico para obtener la estimación final del parámetro.** Este paso calcula el *intervalo de confianza* para ese parámetro.

Interpretar resultados con confianza

Imagina que eres biólogo investigador y estás intentando pescar un pez con una red de mano, y el tamaño de esa red representa el margen de error de un intervalo de confianza. Pongamos que tu nivel de confianza es del 95%. ¿Qué significa eso en realidad? Significa que si lanzas al agua esa red en concreto muchas veces seguidas, cogerás un pez el 95% de las veces. Si has cogido un pez, significa que tu intervalo de confianza era correcto y contiene el parámetro real (en este caso el parámetro está representado por el propio pez).

Pero ¿significa eso que tienes una probabilidad del 95% de coger un pez en cualquier intento? No. ¿Te resulta confuso? Lo cierto es que lo es. Aquí va la explicación: imagina que cierras los ojos y a continuación lanzas la red al agua. En ese momento la probabilidad de coger un pez es del 95%. Ahora recoges la red y vuelves a lanzarla al agua con los ojos todavía cerrados. Después de haber hecho eso, abres los ojos y ves uno de los dos únicos resultados posibles: has cogido un pez o no lo has cogido. La probabilidad ya no tiene nada que ver.

De igual modo, después de haber recopilado los datos y haber calculado el intervalo de confianza, habrás capturado el parámetro poblacional

real o no lo habrás capturado. No podrás decir que estás seguro al 95% de que el parámetro está dentro de tu intervalo concreto. De lo que estás seguro al 95% es del proceso mediante el cual se seleccionan muestras aleatorias y se crean intervalos de confianza (es decir, que a largo plazo el 95% de las veces cogerás un pez).

Sabes que este proceso generará intervalos que capturarán la media poblacional el 95% de las veces. El otro 5% de las veces, los datos recopilados en la muestra tendrán valores anormalmente altos o anormalmente bajos por puro azar, y no representarán a la población. Este 5% mide exclusivamente los errores debidos a la probabilidad aleatoria, y no incluye el sesgo.

El margen de error no significa nada si los datos utilizados en el estudio estaban sesgados o no eran fiables. Sin embargo, no puedes saber tal cosa con sólo mirar los resultados estadísticos de alguien. Mi consejo es que, antes de aceptar el margen de error que te digan, averigües cómo se recopilaron los datos (en los capítulos 16 y 17 hablo de los problemas que entraña la recopilación de datos). Por tanto, antes de creerte un estudio debes hacer algunas preguntas.

Cuestión de amplitud

La *amplitud* de tu intervalo de confianza es dos veces el margen de error. Por ejemplo, supongamos que el margen de error es ±5%. Un intervalo de confianza del 7%, ± 5%, abarca desde el 7% – 5% = 2% hasta el 7% + 5% = 12%. Por lo tanto, la amplitud del intervalo de confianza es de 12% – 2% = 10%. Una manera más sencilla de calcularlo es decir que la amplitud del intervalo de confianza es dos veces el margen de error. En este caso la amplitud del intervalo de confianza es 2 × 5% = 10%.

La amplitud de un intervalo de confianza es la distancia que hay entre el extremo inferior del intervalo (el valor del estadístico menos el margen de error) y el extremo superior del intervalo (el valor del estadístico más el margen de error). Una manera rápida de calcular la amplitud de un intervalo de confianza consiste en multiplicar por dos el margen de error.

Cuando realizas una estimación utilizando un intervalo de confianza, el objetivo final es que ese intervalo sea estrecho, porque eso significa que estás centrándote en el punto donde se encuentra el parámetro. Tener que sumar y restar un margen de error grande hace que tu resultado sea mucho menos preciso.

Entonces, si un margen de error pequeño es algo deseable, ¿uno más pequeño sería todavía mejor? No siempre. Un intervalo de confianza estrecho es algo bueno... hasta cierto punto. Para obtener un intervalo de confianza extremadamente estrecho tienes que realizar un estudio mucho más grande (y más caro), con lo que llega un momento en que el incremento del coste no justifica la pequeña diferencia en cuanto a precisión. La mayoría de la gente se conforma con un margen de error comprendido entre el 2 y el 3% cuando la propia estimación es un porcentaje (por ejemplo el porcentaje de mujeres, de republicanos o de fumadores).

¿Cómo te aseguras de que tu intervalo de confianza será suficientemente estrecho? Desde luego, es importante reflexionar sobre esta cuestión antes de recopilar los datos (después de recopilarlos, la amplitud del intervalo de confianza ya está definida).

Los factores que influyen en la amplitud de un intervalo de confianza son:

- ✔ Nivel de confianza.
- ✔ Tamaño de la muestra.
- ✔ Variabilidad de la población.

Cada uno de estos tres factores influye de manera considerable en la amplitud de un intervalo de confianza. En los siguientes apartados explicaré cada elemento y de qué manera influye.

Elegir un nivel de confianza

A cada intervalo de confianza (y, por tanto, a cada margen de error) le corresponde un porcentaje que representa tu grado de confianza en que los resultados capturarán el parámetro poblacional real, dependiendo de cómo el azar influya en la muestra aleatoria. Este porcentaje se denomina *nivel de confianza*.

El nivel de confianza te ayuda a justificar el resto de los resultados muestrales que podrías haber obtenido cuando estás estimando un parámetro utilizando los datos procedentes de una única muestra. Si quieres justificar el 95% del resto de los resultados posibles, tu nivel de confianza será del 95%.

¿Qué nivel de confianza suelen utilizar los investigadores? Yo he visto niveles de confianza entre el 80 y el 99%, pero el más frecuente es el 95%. De hecho, hay incluso un chiste sobre esta cuestión: ¿por qué a los estadísticos les gusta su trabajo? Porque sólo tienen que acertar el 95 % de las veces (tiene gracia, ¿no? Y creo que los meteorólogos aciertan menos).

La variabilidad en los resultados muestrales se mide por el número de errores estándares. Un *error estándar* es similar a la desviación estándar de un conjunto de datos, sólo que el error estándar se aplica a medias muestrales o porcentajes muestrales que podrías haber obtenido si hubieras tomado otras muestras. (En el capítulo 11 encontrarás información sobre los errores estándares.)

Los errores estándares son los elementos básicos de los intervalos de confianza. Un intervalo de confianza es un estadístico más/menos un margen de error, y el margen de error es el número de errores estándares que necesitas para llegar al nivel de confianza que quieres.

A cada nivel de confianza le corresponde un número de errores estándares que deben sumarse o restarse. Este número de errores estándares es lo que se denomina *valor crítico*. Si estás utilizando una distribución *Z* para encontrar el número de errores estándares (de la forma descrita posteriormente en este mismo capítulo), el valor crítico se llama *valor* z^* (se pronuncia "valor zeta asterisco"). En la tabla 13-1 encontrarás una lista de valores z^* para los intervalos de confianza más comunes.

A medida que aumenta el nivel de confianza, aumenta también el número de errores estándares y, por tanto, el margen de error.

Tabla 13-1. Valores z^* para varios niveles de confianza

Nivel de confianza	*Valor* z^*
80 %	1,28
90 %	1,645 (por convención)
95 %	1,96
98 %	2,33
99 %	2,58

Si quieres un nivel de confianza en los resultados superior al 95%, tendrás que sumar y restar más de dos errores estándares aproximadamente. Por ejemplo, para tener una confianza del 99%, sumas y restas aproximadamente dos errores estándares y medio (2,58, para ser exactos) para obtener el margen de error. Cuanto mayor es el nivel de confianza, más grande es el valor z^* y el margen de error, y más ancho es el intervalo de confianza (suponiendo que todo lo demás permanece igual). Para tener más confianza hay que pagar un precio.

Observa que he dicho "suponiendo que todo lo demás permanece igual". Puedes compensar un incremento del margen de error incrementando también el tamaño de la muestra. Encontrarás más información sobre este tema en el siguiente apartado.

La importancia del tamaño muestral

La relación entre el margen de error y el tamaño muestral es simple: cuando el tamaño muestral aumenta, el margen de error disminuye y el intervalo de confianza se hace más estrecho. Esta relación confirma algo que esperas que sea cierto: cuanta más información (datos) tienes, más precisos serán los resultados (eso, naturalmente, suponiendo que la información sea válida y creíble. En el capítulo 3 explico por qué pueden estar equivocadas las estadísticas).

En todas las fórmulas de margen de error para intervalos de confianza incluidas en este capítulo aparece el tamaño muestral *(n)* en el denominador. Por ejemplo, la fórmula del margen de error para la media muestral, $\pm z^* \frac{\sigma}{\sqrt{n}}$ (que se comenta con más detalle más adelante en este mismo capítulo) tiene una *n* en el denominador de una fracción (igual que la mayoría de las fórmulas de margen de error). Si *n* aumenta, el denominador de esa fracción aumenta, con lo que el valor de la fracción en general se hace más pequeño. Como consecuencia, el margen de error disminuye y el intervalo de confianza se hace más estrecho. Si necesitas un nivel de confianza alto, debes incrementar el valor z^* y, por tanto, el margen de error, con el resultado de un intervalo de confianza más ancho, lo cual no es bueno (ver el apartado anterior). Sin embargo,

este intervalo de confianza más ancho puede compensarse aumentando el tamaño muestral y reduciendo de nuevo el margen de error, con lo que el intervalo de confianza se hace más estrecho.

El aumento del tamaño muestral te permite alcanzar el nivel de confianza que deseas, al tiempo que garantiza un intervalo de confianza estrecho (que es lo que quieres en último término). Es posible incluso determinar el tamaño muestral que necesitas antes de comenzar un estudio: si conoces el margen de error que quieres obtener, basta con definir el tamaño muestral en consecuencia. (Tienes más información en el apartado "Determinar qué tamaño muestral necesitas".)

Si tu estadístico va a ser un porcentaje (por ejemplo el porcentaje de personas que prefieren llevar sandalias en verano), una manera aproximada de determinar el margen de error para un intervalo de confianza del 95% consiste en dividir 1 por la raíz cuadrada de *n* (el tamaño muestral). Puedes probar con diferentes valores de *n* y ver cómo repercuten en el margen de error. Por ejemplo, una encuesta de 100 personas pertenecientes a una población grande tendrá un margen de error aproximado de $\frac{1}{\sqrt{100}}$ =0,1, es decir, ± 10% (por tanto el intervalo de confianza tendrá un ancho del 20%, que es mucho).

No obstante, si preguntas a 1.000 personas, el margen de error disminuye muchísimo, hasta ± 3%, aproximadamente, con lo que la amplitud pasa a ser de tan sólo el 6%. Si la encuesta es de 2.500 personas, el margen de error será de ± 2% (y la amplitud disminuye hasta el 4%). Si piensas en lo grande que es la población (en Estados Unidos, por ejemplo, viven más de 310 millones de personas), este tamaño muestral es bastante pequeño para semejante precisión.

Sin embargo, ten presente que no conviene aumentar en exceso el tamaño muestral, porque llega un punto en que ya no merece la pena. Por ejemplo, si el tamaño muestral pasa de 2.500 a 5.000, la amplitud del intervalo de confianza pasa del 4% al 2 × 1,4=2,8%. Por cada persona más que incorporas a la encuesta el coste aumenta, de manera que preguntar a otras 2.500 personas para estrechar el intervalo un poco más del 1% posiblemente no merezca la pena.

El primer paso en cualquier problema de análisis de datos (y también para valorar críticamente los resultados de otra persona) consiste en cerciorarse de que los datos son válidos. Los resultados estadísticos sólo son válidos si los datos también lo son, de manera que la precisión real depende tanto de la calidad de los datos como del tamaño muestral. Un tamaño muestral grande que tenga mucho sesgo (el asunto está tratado en el capítulo 16) aparentemente tendrá un intervalo de confianza estrecho, pero no significará nada. Es como si compites en tiro con arco y obtienes buenas puntuaciones en todas las rondas, pero al final te das cuenta de que has estado disparando todo el rato a la diana de la persona que tienes al lado; eso te dará una idea de lo equivocado que estás. En el campo de la estadística, sin embargo, no puedes medir el sesgo de forma exacta; tan sólo puedes intentar reducirlo al mínimo diseñando muestras y estudios válidos (para saber cómo tendrás que leer los capítulos 16 y 17).

Tener en cuenta la variabilidad de la población

Uno de los factores que influyen en la variabilidad de los resultados muestrales es el hecho de que la propia población contiene variabilidad. Por ejemplo, en la población formada por las casas que hay en una ciudad grande como Columbus, Ohio, existe mucha variedad, no sólo en cuanto a tipos de casas, sino también en cuanto a tamaños y precios. Además, la variabilidad de los precios de las casas de Columbus probablemente sea mayor que la variabilidad de los precios de las casas de una urbanización en concreto de Columbus.

Eso significa que si tomas una muestra de casas de toda la ciudad de Columbus y calculas el precio medio, el margen de error será más grande que si tomas una muestra de una urbanización concreta de Columbus, incluso si tienes el mismo nivel de confianza y el mismo tamaño muestral.

¿Por qué? Pues porque las casas de toda la ciudad presentan una variabilidad mayor en cuanto a precio, y se supone que tu media muestral cambiaría más de una muestra a otra de lo que cambiaría si tomaras todas las muestras en esa urbanización en concreto, en la que los precios son bastante similares porque todas las casas son parecidas. Por tanto, si eliges casas de toda la ciudad de Columbus tienes que incluir más casas en la muestra para tener el mismo grado de precisión que obtendrías en esa urbanización en concreto.

La desviación estándar de la población se designa con la letra griega σ. Observa que σ aparece en el numerador del error estándar en la fórmula del margen de error para la media muestral: $\pm z^* \frac{\sigma}{\sqrt{n}}$.

Por consiguiente, cuando la desviación estándar (el numerador) aumenta, el error estándar (la fracción entera) también aumenta. Esto se traduce en un margen de error más grande y un intervalo de confianza más ancho. (En el capítulo 11 encontrarás más información sobre el error estándar.) Cuanta más variabilidad haya en la población original, mayor será el margen de error y más ancho será el intervalo de confianza. Este incremento se puede compensar aumentando el tamaño muestral.

Calcular un intervalo de confianzapara una media poblacional

Cuando la característica medida (por ejemplo ingresos, coeficiente intelectual, precio, altura, cantidad o peso) es *numérica*, la mayoría de las personas quieren estimar el valor medio para la población. Para estimar la media poblacional, μ, se utiliza una media muestral, $\bar{x}$, más/menos un margen de error. El resultado se llama *intervalo de confianza para la media poblacional*, μ. Su fórmula depende de si se cumplen ciertas condiciones. Divido las condiciones en dos casos, ilustrados en los siguientes apartados.

Caso 1: Se conoce la desviación estándar de la población

En el caso 1, la desviación estándar de la población es conocida. La fórmula de un intervalo de confianza (IC) para una media poblacional en este caso es $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$, en la que $\bar{x}$ es la media muestral, σ es la desviación estándar de la población, *n* es el tamaño muestral, y z^* representa el valor z^* adecuado de la distribución normal estándar para el nivel de confianza que deseas (en la tabla 13-1 encontrarás los valores de z^* para varios niveles de confianza).

En este caso los datos deben provenir de una distribución normal o bien, si no es así, *n* debe ser suficientemente grande (al menos 30) como para poder aplicar el teorema del límite central (ver el capítulo 11), y así poder utilizar los valores z^* en la fórmula.

A fin de calcular un IC para la media poblacional bajo las condiciones del caso 1, haz lo siguiente:

1. **Determina el nivel de confianza y encuentra el valor z^* adecuado.**

   Consulta la tabla 13-1.

2. **Encuentra la media muestral *($\bar{x}$)* para el tamaño muestral *(n)*.**

   Nota: se supone que la desviación estándar de la población es un valor conocido, σ.

3. **Multiplica z^* por σ y divide el resultado por la raíz cuadrada de *n*.**

   Este cálculo proporciona el margen de error.

4. **El IC es ± $\bar{x}$ el margen de error.**

El extremo inferior del IC es $\bar{x}$ menos el margen de error, y el extremo superior del IC es $\bar{x}$ más el margen de error.

Por ejemplo, imagina que trabajas para el Departamento de Recursos Naturales y quieres estimar, con un intervalo de confianza del 95%, la longitud media de los alevines de trucha en una piscifactoría.

1. Como quieres un intervalo de confianza del 95%, tu valor z^* es 1,96.

2. Pongamos que coges una muestra de 100 alevines y determinas que la longitud media es 7,5 centímetros, y supongamos que la desviación estándar es 2,3 centímetros. Esto significa que $\bar{x}$ = 7,5, σ = 2,3 y *n* = 100.

3. Multiplica 1,96 por 2,3 y divide por la raíz cuadrada de 100 (que es 10). El margen de error es, pues, ± 1,96 × (2,3 / 10) = 1,96 × 0,23 = 0,45 centímetros.

4. Tu intervalo de confianza del 95% para la longitud media de los alevines de lucioperca en esa piscifactoría es 7,5 centímetros ±0,45 centímetros (el extremo inferior del intervalo es 7,5 – 0,45 = 7,05 centímetros, y el extremo superior es 7,5 + 0,45 = 7,95 centímetros).

Después de calcular un intervalo de confianza, asegúrate de interpretarlo con palabras comprensibles para alguien que no sepa de estadística. Es decir, procura exponer los resultados refiriéndote a lo que la persona del problema intenta averiguar. Esto es lo que en estadística se llama interpretar los resultados "en el contexto del problema". En este ejemplo puedes decir lo siguiente: "A partir de los datos de mi muestra, y con un

nivel de confianza del 95%, la longitud media de los alevines de trucha de esta piscifactoría está entre 7,05 y 7,95 centímetros".

Caso 2: Se desconoce la desviación estándar de la población y/o *n* es pequeño

En muchas situaciones σ es desconocida, de manera que se estima con la desviación estándar de la muestra, *s*; o el tamaño muestral es pequeño (inferior a 30), y no puedes saber si los datos proceden de una distribución normal (en este último caso no se puede aplicar el teorema del límite central; ver el capítulo 11). En ninguna de las dos situaciones puedes utilizar un valor z^* de la distribución normal estándar (distribución *Z*) como valor crítico; tienes que utilizar un valor crítico más grande porque desconoces σ o porque tienes menos datos.

La fórmula del intervalo de confianza para una media poblacional en el caso 2 es $\bar{x} \pm t^*_{n-1} \frac{s}{\sqrt{n}}$, en la que t^*_{n-1} es el valor t^* crítico de la distribución *t* que tiene *n*–1 grados de libertad (donde *n* es el tamaño muestral). Los valores t^* para los niveles de confianza habituales se encuentran mirando en la última fila de la tabla *t* (que encontrarás en el apéndice). En el capítulo 10 encontrarás información detallada sobre la distribución *t* y sobre cómo utilizar la tabla *t*.

La distribución *t* tiene una forma similar a la distribución *Z*, salvo que es más plana y sus valores están más dispersos. Para valores pequeños de *n* y un determinado nivel de confianza, los valores críticos en la distribución *t* son más grandes que en la distribución *Z*, de manera que cuando utilices los valores críticos de la distribución *t* el margen de error para tu intervalo de confianza será mayor. Conforme los valores de *n* se hacen más grandes, los valores t^* se acercan más a los valores z^*. (En el capítulo 10 encontrarás información detallada sobre la distribución *t* y sus relaciones con la distribución *Z*.)

En el ejemplo de la piscifactoría del caso 1, el tamaño muestral era 10 en lugar de 100, y todo lo demás era igual. El valor t^* en este caso viene de una distribución *t* con 10–1=9 grados de libertad. Este valor t^* se encuentra mirando en la tabla *t* (incluida en el apéndice). Primero miras en la última fila, donde están los niveles de confianza, y localizas el nivel de confianza del 95%, que señala la columna que necesitas. A continuación buscas la fila correspondiente a *gl*=9. En la intersección de esa fila y esa columna encuentras t^*=2,262. Éste es el valor t^* correspondiente a un intervalo de confianza del 95% para la media con un tamaño muestral de 10 (observa que es más grande que el valor z^* de 1,96 encontrado en la tabla 13-1). Si aplicas la fórmula para calcular el intervalo de confianza, obtienes $7{,}5 \pm 2{,}262 \frac{2{,}3}{\sqrt{10}} = 7{,}50 \pm 1{,}645$, es decir, entre 5,86 y 9,15 centímetros.

(En el capítulo 10 encontrarás información detallada sobre la distribución *t* y sobre cómo utilizar la tabla *t*.)

Observa que este intervalo de confianza es más ancho que el calculado para *n*=100. Además de tener un valor crítico más grande (t^* en lugar de z^*), el tamaño de la muestra es mucho menor, con lo que el margen de error aumenta porque *n* está en el denominador.

En el supuesto de que tengas que utilizar *s* porque no conoces σ, el intervalo de confianza también será más ancho. También suele ocurrir que σ no se conoce y el tamaño muestral es pequeño, en cuyo caso el intervalo de confianza también es más ancho.

Determinar qué tamaño muestral necesitas

El margen de error de un intervalo de confianza está relacionado con el tamaño muestral (puedes releer el apartado anterior "La importancia del tamaño muestral"): cuanto más grande es la muestra, más pequeño es el margen de error. Si lo piensas al revés, verás que para conseguir un margen de error pequeño (lo que siempre quieres) necesitas un tamaño muestral grande. Imagina que estás preparándote para hacer tu propia encuesta con el fin de estimar una media poblacional. ¿No estaría bien saber de antemano qué tamaño muestral necesitas para conseguir el margen de error que quieres? Pensar las cosas con antelación te ahorrará tiempo y dinero y te proporcionará resultados con un margen de error razonable, para que luego no tengas sorpresas.

La fórmula con la que averiguar el tamaño muestral necesario para obtener el margen de error (ME) que deseas cuando estás elaborando un intervalo de confianza para μ es $\geq \left(\frac{z^* \sigma}{MOE} \right)^2$. El tamaño muestral debe redondearse siempre hacia arriba, sea cual sea el valor decimal obtenido (por ejemplo, si los cálculos arrojan un tamaño muestral de 126,2 personas, no puedes coger el 0,2 de una persona. Necesitas la persona completa, así que la incluyes redondeando a 127).

En esta fórmula, ME es el número que representa el margen de error deseado, y z^* es el valor z^* correspondiente al nivel de confianza que buscas (según la tabla 13-1; generalmente se utiliza el valor 1,96 para un intervalo de confianza del 95%). Si desconoces la desviación estándar de la población, σ, puedes utilizar el valor aproximado que correspondería a la situación más desfavorable posible o bien puedes hacer un estudio

preliminar (un pequeño estudio de prueba), encontrar la desviación estándar de los datos muestrales *(s)* y utilizar ese número. Esto puede ser arriesgado si el tamaño muestral es muy pequeño, ya que es menos probable que refleje a la población entera. Por tanto, si haces un estudio de prueba, procura que sea lo más grande posible. De otro modo, mejor que seas conservador en tu estimación de σ.

Generalmente merece la pena dedicar algo de tiempo y esfuerzo a un pequeño estudio preliminar. Además de obtener una estimación de σ que te ayudará a elegir el tamaño muestral adecuado, puede que descubras posibles problemas en el proceso de recopilación de datos.

En este capítulo únicamente pongo una fórmula para calcular el tamaño muestral: la que corresponde a un intervalo de confianza para una media poblacional (no obstante, puedes utilizar la fórmula aproximada que menciono en el apartado anterior "La importancia del tamaño muestral" cuando se trate de proporciones).

Veamos un ejemplo en el que necesitas calcular *n* para estimar una media poblacional: imagina que quieres estimar el número medio de canciones que los universitarios tienen almacenadas en sus reproductores portátiles. Quieres que el margen de error no sea superior a ± 20 canciones, y quieres un intervalo de confianza del 95%. ¿Cuántos universitarios deben componer tu muestra?

Como quieres un IC del 95%, z^* es 1,96 (según la tabla 13-1); y sabes que el ME deseado es 20. Ahora necesitas un número para la desviación estándar de la población, σ. Este número es desconocido, así que haces un estudio preliminar con 35 universitarios y ves que la desviación estándar *(s)* para la muestra es de 148 canciones: éste será el número que utilices como sustituto de σ. Aplicando la fórmula, calculas que el tamaño muestral necesario es $n \geq \left(\frac{1{,}96(148)}{20}\right)^2 = (14{,}504)^2 = 210{,}37$, cifra que redondeas a 211 universitarios (al calcular *n* hay que redondear siempre al alza). Por lo tanto, debes tomar una muestra aleatoria de al menos 211 universitarios para que el margen de error en el número de canciones almacenadas no sea superior a 20. Por eso en la fórmula aparece el signo ≥, mayor o igual que.

Cuando calcules el tamaño muestral debes redondear siempre al siguiente número entero, con independencia del valor decimal que hayas obtenido (por ejemplo 0,37). Esto es así porque el margen de error no debe ser mayor que el valor definido. Si redondearas a la baja al obtener un valor decimal inferior a 0,50 (como sueles hacer en otros cálculos matemáticos), el ME sería un poco mayor que lo deseado.

Si te estás preguntando de dónde sale esta fórmula para el tamaño muestral, es el resultado de algunos cálculos matemáticos. Coges la fórmula del margen de error (donde aparece *n*), sustituyes el resto de las variables de la fórmula por los números que sacas del problema, igualas al ME deseado y despejas para *n*.

Determinar el intervalo de confianza para una proporción poblacional

Si la característica medida es categórica —por ejemplo la opinión sobre una determinada cuestión (a favor, en contra o indiferente), el sexo, la filiación política o el tipo de comportamiento (llevar/no llevar el cinturón de seguridad puesto al conducir)—, generalmente se quiere estimar la proporción (o porcentaje) de personas de la población que pertenecen a cierta categoría de interés. Piensa, por ejemplo, en el porcentaje de personas que están a favor de la semana laboral de cuatro días, el porcentaje de republicanos que acudieron a las urnas en las pasadas elecciones en Estados Unidos o la proporción de conductores que no llevan puesto el cinturón de seguridad. En todos estos casos se trata de estimar una proporción poblacional, *p*, utilizando una proporción muestral, $\hat{p}$, más/menos un margen de error. El resultado se llama *intervalo de confianza para la proporción poblacional*, *p*.

La fórmula con la que se halla el IC para una proporción poblacional es $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ en la que $\hat{p}$ es la proporción poblacional, *n* es el tamaño muestral, y z^* es el valor de la distribución normal estándar que resulta adecuado para el nivel de confianza que deseas. En la tabla 13-1 encontrarás los valores z^* correspondientes a varios niveles de confianza.

Para calcular un IC para la proporción poblacional, haz lo siguiente:

1. **Determina el nivel de confianza y encuentra el valor z^* adecuado.**

   Encontrarás varios valores z^* en la tabla 13-1.

2. **Encuentra la proporción muestral, $\hat{p}$, dividiendo el número de personas de la muestra que poseen la característica de interés**

por el tamaño muestral *(n)*. Ten en cuenta que el resultado debe ser un valor decimal entre 0 y 1.

3. **Multiplica $\hat{p}$ (1 - $\hat{p}$) y luego divide el resultado por *n*.**

4. **Calcula la raíz cuadrada del resultado del paso 3.**

5. **Multiplica ese valor por z^*.**

   Este cálculo proporciona el margen de error.

6. **El IC es $\hat{p}$ ± el margen de error. El extremo inferior del IC es $\hat{p}$– el margen de error, y el extremo superior del IC es $\hat{p}$+ el margen de error.**

La fórmula indicada en el ejemplo anterior, con la que se determina el IC para *p*, se utiliza con la condición de que el tamaño muestral sea suficientemente grande como para poder aplicar el teorema del límite central y así poder utilizar un valor z^* (un asunto del capítulo 11), lo cual ocurre en aquellos casos en que se estima una proporción a partir de una encuesta a gran escala (esto se trata en el capítulo 9). Para tamaños muestrales pequeños, los intervalos de confianza para la proporción están fuera del alcance de un curso de introducción a la estadística.

Por ejemplo, imagina que quieres estimar el porcentaje de veces que te tropiezas con un semáforo en rojo en un determinado cruce.

1. Como quieres un intervalo de confianza del 95%, tu valor z^* es 1,96.

2. Coges una muestra aleatoria de 100 viajes diferentes que pasen por ese cruce y compruebas que te encuentras con el semáforo rojo en 53 ocasiones, de manera que $\hat{p}$ = 53 / 100 = 0,53.

3. Encuentra $\hat{p}$ **(1 -** $\hat{p}$**)** = 0,53* (1 – 0,53) = 0,2491 / 100 = 0,002491.

4. Calcula la raíz cuadrada para obtener 0,0499.

   El margen de error es, pues, ± 1,96 × (0,0499) = 0,0978, es decir, el 9,78%.

5. El intervalo de confianza del 95% para el porcentaje de veces que te encuentras con el semáforo en rojo en ese cruce en concreto es 0,53 (o el 53%) más/menos 0,0978 (redondeamos a 0,10, que es el 10%). El extremo inferior del intervalo es 0,53 − 0,10 = 0,43, o sea, el 43%; el extremo superior es 0,53 + 0,10 = 0,63, el 63%.

Para interpretar estos resultados en el contexto del problema, puedes decir que, basándote en los resultados de tu muestra y con un nivel de confianza del 95%, el porcentaje de veces que te encuentras con un semáforo en rojo en ese cruce está entre el 43 y el 63%.

Cuando hagas cálculos con porcentajes muestrales, utiliza siempre la forma decimal. Una vez que hayas terminado los cálculos, convierte a porcentajes multiplicando por 100. Para evitar el error de redondeo, conserva siempre dos decimales por lo menos.

Crear un intervalo de confianza para la diferencia entre dos medias

Muchas encuestas y estudios tienen el propósito de comparar dos poblaciones, por ejemplo hombres frente a mujeres, ingresos bajos frente a ingresos elevados, o republicanos frente a demócratas. Cuando la característica comparada es numérica (por ejemplo altura, peso o ingresos), el objeto de interés es la magnitud de la diferencia entre las medias de las dos poblaciones.

Por ejemplo, puede que quieras comparar la diferencia de la edad que tienen de media los republicanos y los demócratas, o la diferencia de los ingresos medios de hombres y mujeres. Para estimar la diferencia entre dos medias poblacionales, $\mu_1 - \mu_2$, hay que tomar una muestra de cada población (llamémoslas muestra 1 y muestra 2) y coger la diferencia entre ambas medias muestrales, $\bar{x}_1 - \bar{x}_2$, ± un margen de error. El resultado es un *intervalo de confianza para la diferencia entre dos medias poblacionales*, $\mu_1 - \mu_2$. La fórmula para hallar el IC varía en función de ciertas condiciones, como hemos visto en los apartados anteriores. Tenemos el caso 1 y el caso 2.

Caso 1. Se conocen las desviaciones estándares de la población

El caso 1 presupone que las desviaciones estándares de las dos poblaciones son conocidas. La fórmula con la que se calcula el IC para la diferencia entre dos medias poblacionales es $\bar{x}_1 - \bar{x}_2 \pm z^* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$, en la que $\bar{x}_1$ y n_1 son la media y el tamaño de la primera muestra (la desviación estándar de la primera población, σ_1, nos la dan), y $\bar{x}_2$ y n_2 son la media y el tamaño de la segunda muestra (la desviación estándar de la segunda población, σ_2, también nos la dan). Aquí z^* es el valor de la distribución normal estándar que es adecuado para el nivel de confianza que deseas (en la tabla 13-1 encontrarás valores de z^* para varios niveles de confianza).

Para calcular un IC para la diferencia entre dos medias poblaciones, haz lo siguiente:

1. **Determina el nivel de confianza y encuentra el valor z^* adecuado.**

   Consulta la tabla 13-1.

2. **Identifica $\bar{x}_1$, n_1 y σ_1, así como , $\bar{x}_2$ n_2 y σ_2.**

3. **Calcula la diferencia, ($\bar{x}_1 - \bar{x}_2$), entre las medias muestrales.**

4. **Eleva al cuadrado σ_1 y divide el resultado por n_1; eleva al cuadrado σ_2 y divide el resultado por n_2. Suma los resultados y calcula la raíz cuadrada.**

5. **Multiplica la respuesta del paso 4 por z^*.** Este cálculo proporciona el margen de error.

6. **El IC es $\bar{x}_1 - \bar{x}_2$ ± el margen de error.**

El extremo inferior del IC es $\bar{x}_1 - \bar{x}_2$ - margen de error, y el extremo superior del IC es $\bar{x}_1 - \bar{x}_2$ + margen de error.

Pongamos que quieres estimar con un nivel de confianza del 95% cuál es la diferencia entre la longitud media de las espigas de dos variedades distintas de trigo (permitiendo que crezcan el mismo número de días bajo las mismas condiciones). Llamaremos a las dos variedades A y B. Por investigaciones anteriores sabes que las desviaciones estándares de población para las variedades A y B son 0,35 centímetros y 0,45 centímetros, respectivamente.

1. Como quieres un intervalo de confianza del 95%, tu valor z^* es 1,96.

2. Supongamos que la muestra aleatoria de 100 espigas de la variedad A tiene una longitud media de 8,5 centímetros, y la muestra aleatoria de 110 espigas de la variedad B tiene una longitud media de 7,5 centímetros. Así pues, tienes la información siguiente: $\bar{x}_1$ =8,5, σ_1 = 0,35, n_1, n_1=100, $\bar{x}_2$ = 7,5, σ_2 = 0,45 y n_2 = 110.

3. Según el paso 3 hay que calcular la diferencia entre las medias muestrales, $\bar{x}_1 - \bar{x}_2$, que es 8,5 − 7,5 = +1 centímetro. La media de la variedad A menos la media de la variedad B es un valor positivo, de manera que la variedad A es la variedad más grande de las dos, a partir de los datos de esta muestra. Pero ¿esta diferencia es suficiente para extrapolar el resultado a toda la población? Eso es lo que vas a decidir con la ayuda de este intervalo de confianza.

4. Calculas el cuadrado de σ_1 (0,35), que es 0,1225; divides por 100 y obtienes 0,0012. Calculas el cuadrado de σ_2 (0,45) y divides por 110, con lo que resulta 0,2025 / 110 = 0,0018. La suma es 0,0012 + 0,0018 = 0,0030; la raíz cuadrada es 0,0554 centímetros (sin redondeo).

5. Multiplicas 1,96 × 0,0554; te sale 0,1085 centímetros, el margen de error.

6. El intervalo de confianza del 95% para la diferencia entre las longitudes medidas de estas dos variedades de trigo es 1 centímetro ± 0,1085 centímetros (el extremo inferior del intervalo es 1 − 0,1085 = 0,8915 centímetros, y el extremo superior es 1 + 0,1085 = 1,1085 centímetros). Observa que todos los valores de este intervalo son positivos. Por tanto, estimas que la variedad A es más larga que la variedad B, según los datos que manejas.

Para interpretar estos resultados en el contexto del problema, puedes decir que, basándote en los resultados de tu muestra y con un nivel de confianza del 95%, la variedad A es más larga, en promedio, que la variedad B, y la diferencia está entre 0,8915 y 1,1085 centímetros.

Ten presente que $\bar{x}_1 - \bar{x}_2$ podría dar como resultado un valor negativo. Por ejemplo, si hubieras intercambiado las dos variedades de trigo, la diferencia habría sido –1. En ese caso habrías dicho que la variedad B era, en promedio, un centímetro más corta que la variedad A (la misma conclusión expresada de manera diferente).

Si quieres evitar valores negativos para la diferencia entre medias muestrales, pon primero el grupo que tenga la media muestral más grande. De este modo la diferencia será positiva (es lo que hago yo).

Caso 2. Se desconocen las desviaciones estándares de la población y/o los tamaños muestrales son pequeños

Es posible que desconozcas σ_1 y σ_2, en cuyo caso las estimas con las desviaciones estándares de las muestras, $s1_1$ y $s1_2$, y/o que los tamaños muestrales sean pequeños (inferiores a 30) y no estés seguro de si los datos proceden de una distribución normal.

En el caso 2, el intervalo de confianza para la diferencia entre dos medias poblacionales es $(\bar{x}_1 - \bar{x}_2) \pm t^*_{n_1+n_2-2}\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$, en la que t^* es el valor crítico de la distribución *t* que tiene $n_1 + n_2 - 2$ grados de libertad; n_1 y n_2 son los dos tamaños muestrales, respectivamente; y s_1 y s_2 son las dos desviaciones estándares de las muestras. Este valor t^* se encuentra en la tabla *t* (está en el apéndice), buscando la intersección entre la fila correspondiente a $gl = n_1 + n_2 - 2$ y la columna correspondiente al nivel de confianza que necesites, según indique la última fila de la tabla (mira el capítulo 10). Aquí suponemos que las desviaciones estándares de las poblaciones son similares. De lo contrario, se introduce la modificación correspondiente utilizando el error estándar y los grados de libertad. (Consulta el final del apartado sobre comparación de dos medias, en el capítulo 15.)

Volviendo al ejemplo del caso 1, imagina que las espigas de las dos variedades de maíz, la variedad A (grupo 1) y la variedad B (grupo 2), tienen la misma longitud media que antes: $\bar{x}_1$=8,5 y $\bar{x}_2$ = 7,5 centímetros. Esta vez, sin embargo, no conoces las desviaciones estándares de las poblaciones, de manera que utilizas en su lugar las desviaciones estándares de las muestras (pongamos que son s_1 = 0,40 y s_2 = 0,50 centímetros, respectivamente). Supongamos que los tamaños muestrales, n_1 y n_2, son solamente 15 en este caso.

Para calcular el IC, primero hay que encontrar el valor t^* de la distribución *t* que tiene (15 + 15 – 2) = 28 grados de libertad (suponemos que el nivel de confianza continúa siendo del 95%). En la tabla *t* (la tienes en el apéndice), busca la fila correspondiente a 28 grados de libertad y la columna que representa un nivel de confianza del 95% (mira las etiquetas que hay en la última fila de la tabla); encuentra la intersección de fila y columna y verás que t^*_{28} = 2,048. Utilizando el resto de la información que te dan, el intervalo de confianza para la diferencia entre las longitudes medias de las espigas de las dos variedades es $(8{,}5-7{,}5) \pm 2{,}048\sqrt{\frac{(15-1)(0{,}4)^2+(15-1)(0{,}5)^2}{15+15-2}}$ =1 ± 2,048 (0,45) = 1 ± 0,9273 centímetros.

Eso significa que el IC del 95% para la diferencia entre las longitudes medias de las espigas de estas dos variedades de trigo en esta situación es (0,0727, 1,9273) centímetros, y las espigas de la variedad A son las más largas. (Advierte que este IC es más ancho que el obtenido para el caso 1, como era de esperar.)

Estimar la diferencia entre dos proporciones

Cuando la característica de los dos grupos que se comparan es *categórica*, por ejemplo la opinión sobre una determinada cuestión (a favor/en contra), lo que se quiere plasmar son las diferencias entre las dos proporciones poblacionales, por ejemplo la diferencia entre la proporción de mujeres que están a favor de la semana de cuatro días laborables y la proporción de hombres que están a favor de la semana de cuatro días laborables. ¿Y esto cómo se hace?

Para estimar la diferencia entre dos proporciones poblacionales, $p_1 - p_2$, hay que tomar una muestra de cada población y coger la diferencia entre ambas proporciones muestrales, $\hat{p}_1 - \hat{p}_2$, ± un margen de error. El resultado es un intervalo de confianza para la diferencia entre dos proporciones poblacionales, $p_1 - p_2$.

La fórmula con la que se calcula el IC para la diferencia entre dos proporciones poblacionales es $(\hat{p}_1 - \hat{p}_2) \pm z^*\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$, en la que $\hat{p}_1$ y n_1 son la proporción muestral y el tamaño muestral de la primera muestra, y $\hat{p}_2$ y n_2 son la proporción muestral y el tamaño muestral de la segunda muestra. z^* es el valor de la distribución normal estándar que es adecuado para el nivel de confianza que deseas (encontrarás valores z^* en la tabla 13-1).

Para calcular el IC correspondiente a la diferencia entre dos proporciones poblaciones, haz lo siguiente:

1. **Determina el nivel de confianza y encuentra el valor z^* adecuado.**

   Consulta la tabla 13-1.

2. **Encuentra la proporción muestral $\hat{p}_1$ de la primera muestra, para lo cual debes coger el número total de elementos de la primera muestra que pertenecen a la categoría de interés y dividirlo por el tamaño muestral, n_1. De forma similar, encuentra la proporción muestral $\hat{p}_2$ de la segunda muestra.**

3. **Calcula la diferencia entre las proporciones muestrales, $\hat{p}_1 - \hat{p}_2$.**

4. **Calcula $(1 - \hat{p})$ y luego divide el resultado por n_1. Calcula $\hat{p}_1$ $(1 - \hat{p}_2)$ y luego divide el resultado por n_2. Suma estos dos resultados y calcula la raíz cuadrada.**

5. **Multiplica z^* por el resultado del paso 4.**

   Este cálculo proporciona el margen de error.

6. **El IC es $\hat{p}_{1} - \hat{p}_{>sub>2}$ ± el margen de error obtenido en el paso 5.**

El extremo inferior del IC es $\hat{p}_1 - \hat{p}_2$ – el margen de error, y el extremo superior del IC es $\hat{p}_1 - \hat{p}_2$ + más el margen de error.

La fórmula aquí indicada, con la que se determina el IC para $p_1 - p_2$, se utiliza con la condición de que los dos tamaños muestrales sean suficientemente grandes como para poder aplicar el teorema del límite central y así poder utilizar un valor z^* (ver el capítulo 11), lo cual ocurre cuando se estiman proporciones utilizando encuestas a gran escala, por ejemplo. Para tamaños muestrales pequeños, los intervalos de confianza

están fuera del alcance de un curso de introducción a la estadística.

Imagina que trabajas para la Cámara de Comercio de Las Vegas y quieres estimar con un nivel de confianza del 95% cuál es la diferencia entre el porcentaje de mujeres que alguna vez han ido a ver a un imitador de Elvis y el porcentaje de hombres que alguna vez han ido a ver a un imitador de Elvis, con el propósito de determinar cuál es la mejor estrategia de marketing para la oferta de espectáculos de la ciudad.

1. Como quieres un intervalo de confianza del 95%, tu valor z^* es 1,96.

2. Supón que en tu muestra aleatoria de 100 mujeres hay 53 mujeres que en alguna ocasión han ido a ver el espectáculo de un imitador de Elvis, de manera que $\hat{p}_1$ es 53 / 100 =0,53. Supón también que en tu muestra aleatoria de 110 hombres hay 37 hombres que en alguna ocasión han ido a ver el espectáculo de un imitador de Elvis, de manera que $\hat{p}_2$ es 37 / 110 =0,34.

3. La diferencia entre estas proporciones muestrales (mujeres-hombres) es 0,53 – 0,34 =0,19.

4. Calculas 0,53 × (1 − 0,53) y lo divides por 100, con lo que resulta 0,2491 / 100 = 0,0025. Ahora calculas 0,34 × (1 − 0,34) y lo divides por 110, con lo que resulta 0,2244 / 110 = 0,0020. Si sumas estos dos resultados, te da 0,0025 + 0,0020 = 0,0045, cuya raíz cuadrada es 0,0671. 5. 1,96 × 0,0671 = 0,13, es decir, el 13%, que es el margen de error.

6. El intervalo de confianza del 95% para la diferencia entre el porcentaje de mujeres que han visto a un imitador de Elvis y el porcentaje de hombres que han visto a un imitador de Elvis es 0,19, es decir, el 19% (valor obtenido en el paso 3), ± 13%. El extremo inferior del intervalo es 0,19 – 0,13 = 0,06, o sea, el 6%; el extremo superior es 0,19 + 0,13 = 0,32, el 32%.

Para interpretar estos resultados en el contexto del problema, puedes decir que, basándote en los resultados de tu muestra y con un nivel de confianza del 95%, hay un porcentaje mayor de mujeres que de hombres que hayan visto alguna vez a un imitador de Elvis, y la diferencia entre ambos porcentajes está entre el 6 y el 32%.

Estoy pensando que algunos hombres no admitirían nunca haber asistido al espectáculo de un imitador de Elvis (aunque probablemente ellos mismos lo hayan imitado alguna vez en un karaoke). Esto podría generar cierto sesgo en los resultados (la última vez que estuve en Las Vegas creo que vi al verdadero Elvis; iba conduciendo un taxi camino del aeropuerto).

Ten presente que $\hat{p}_1 – \hat{p}_2$ podría dar como resultado un valor negativo. Por ejemplo, si hubieras intercambiado los hombres y las mujeres, la diferencia habría sido –0,19. No es que pase nada, pero puedes evitar diferencias negativas en las proporciones muestrales poniendo primero el grupo que tenga la proporción muestral más grande (en este caso, las mujeres).

Detectar intervalos de confianza engañosos

Cuando el ME es pequeño, en términos relativos, tiendes a creer que esos intervalos de confianza proporcionan estimaciones precisas y fiables de sus parámetros. Sin embargo, no siempre es así.

No todas las estimaciones son tan precisas y fiables como las fuentes quieren hacerte creer. Por ejemplo, una encuesta de un sitio web basada en 20.000 visitas puede que tenga un ME pequeño según la fórmula, pero el ME no significa nada si la encuesta se hace exclusivamente a las personas que entren en ese sitio web.

Dicho de otro modo, la muestra dista mucho de ser una muestra aleatoria (donde todas las muestras de igual tamaño seleccionadas entre la población tienen la misma probabilidad de ser elegidas para participar). No obstante, es frecuente que se publiquen este tipo de resultados, junto con los márgenes de error que hacen que el estudio parezca verdaderamente científico. ¡Mucho ojo con estos resultados fraudulentos! (En el capítulo 12 encontrarás más información sobre los límites del ME.)

Antes de tomar una decisión basada en la estimación de otra persona, haz lo siguiente:

- ✔ Entérate de dónde proviene la estadística; debe ser el resultado de un proceso científico que genere datos fiables, no sesgados y exactos.
- ✔ Averigua cuál es el margen de error. Si no lo mencionan, pídeselo a la fuente original.
- ✔ Recuerda que si la estadística no es fiable o tiene sesgo, el margen de error no significa nada.

(En el capítulo 16 explico cómo evaluar los datos de una encuesta y en el capítulo 17 expongo las condiciones que deben cumplirse para que los

datos de los experimentos se consideren válidos.)

Capítulo 14

Afirmaciones, contrastes y conclusiones

En este capítulo

- Verificar las afirmaciones de otras personas
- Utilizar contrastes de hipótesis para ponderar las pruebas y tomar decisiones
- Comprender que tus conclusiones pueden ser erróneas

La gente continuamente hace afirmaciones relacionadas con estadísticas, y los medios de comunicación no son una excepción:

- ✔ El 25% de las mujeres estadounidenses tienen varices (uf, hay cosas que casi es mejor no saber, ¿verdad?).
- ✔ El consumo de tabaco en Estados Unidos continúa bajando. El porcentaje de fumadores disminuye aproximadamente el 2% cada año desde hace diez.
- ✔ Un bebé de seis meses duerme una media de catorce a quince horas diarias (sí, ya, y yo me lo creo).
- ✔ Las tartas de masa preparada de la marca X están listas en sólo cinco minutos.

Estamos en la era de la información (y del dinero), y es importantísimo poder respaldar lo que uno dice. Si una empresa afirma que sus productos son mejores que los productos de la marca líder, más vale que pueda probarlo o podría caerle una buena demanda. Para que la Food and Drug Administration (FDA) apruebe un medicamento, debe tener pruebas claras de que funciona y no produce efectos secundarios potencialmente mortales. Los fabricantes deben cerciorarse de que sus productos se fabrican con arreglo a las especificaciones si quieren evitar que sean retiradas, quejas de los clientes y pérdidas del volumen de ventas.

Aunque muchas afirmaciones están respaldadas por investigaciones científicas serias (que manejan estadísticas fiables), hay otras que no. En este capítulo aprenderás a utilizar la estadística para averiguar si una afirmación es válida y conocerás el proceso que los investigadores deberían seguir para validar sus afirmaciones.

Un *contraste de hipótesis* es un procedimiento estadístico diseñado para verificar una afirmación. Antes de entrar en detalles, quiero darte una visión general de un contraste de hipótesis mostrándote los pasos que lo componen. Estos pasos se comentan en los siguientes apartados.

1. **Formular la hipótesis nula y la hipótesis alternativa.**
2. **Recopilar datos válidos utilizando un estudio bien diseñado** (como se explica en los capítulos 16 y 17).
3. **Calcular el estadístico de contraste a partir de los datos.**
4. **Hallar el valor *p* para el estadístico de contraste.**
5. **Decidir si se rechaza o no se rechaza H_0 en función del valor *p*.**
6. **Comprender que la conclusión alcanzada podría ser errónea por puro azar.**

Formular las hipótesis

Por regla general, la afirmación contrastada se refiere a un *parámetro poblacional* (un número que caracteriza a toda la población). Como los parámetros suelen ser cantidades desconocidas, todo el mundo quiere hacer afirmaciones sobre sus valores posibles. Por ejemplo, decir que el 25% (o el 0,25) de las mujeres estadounidenses tienen varices es una afirmación referida a la proporción (ése es el *parámetro*) de las mujeres estadounidenses (ésa es la *población*) que tienen varices (ésa es la *variable*, tener o no varices).

Los investigadores a menudo cuestionan afirmaciones referidas a parámetros poblacionales. Por ejemplo, puedes plantear la hipótesis de que en realidad la proporción de mujeres estadounidenses que tienen varices es inferior a 0,25, según tus observaciones. O también puedes plantear la hipótesis de que, debido a que los zapatos de tacón están de moda, la proporción es superior a 0,25. O bien, si sencillamente cuestionas que la proporción real sea 0,25, tu hipótesis alternativa será: "No, no es 0,25".

Definir la hipótesis nula

Todo contraste de hipótesis consta de dos afirmaciones opuestas, o hipótesis, referidas a un parámetro poblacional. La primera hipótesis se denomina *hipótesis nula*, y se designa como H_0. La hipótesis nula siempre dice que el parámetro poblacional es igual al valor propuesto. Por ejemplo, si la afirmación es que el tiempo medio que tarda en hornearse una tarta de masa preparada de la marca X es 5 minutos, la notación estadística abreviada para la hipótesis nula en este caso sería la siguiente: H_0: $\mu=5$ (es decir, la media poblacional es 5 minutos).

La hipótesis nula siempre lleva el signo de igual; H_0 no puede contener el signo≤ni el signo ≥. No es que quiera escurrir el bulto, pero la razón de que esto sea así escapa al alcance de este libro. Digamos simplemente que no me pagarías para que te lo explicara.

¿Cuál es la alternativa?

Antes de llevar a cabo el contraste de hipótesis tienes que poner dos hipótesis posibles sobre la mesa, una de las cuales es la hipótesis nula. Pero si la hipótesis nula se rechaza (por haberse encontrado suficientes pruebas en su contra), ¿cuál será la alternativa? En realidad existen tres posibilidades para la hipótesis alternativa, designada como H_a. Pongo a continuación sus notaciones abreviadas en el contexto del ejemplo de las tartas:

- ✔ El parámetro poblacional es *distinto* del valor propuesto (H_a: $\mu \neq 5$).
- ✔ El parámetro poblacional es *mayor que* el valor propuesto (H_a: $\mu>5$).
- ✔ El parámetro poblacional es *menor que* el valor propuesto (H_a: $\mu<5$).

La hipótesis alternativa que elijas para el contraste de hipótesis depende de la conclusión a la que quieras llegar, si es que reúnes pruebas suficientes para refutar la hipótesis nula (la afirmación).

Por ejemplo, si quieres comprobar si una empresa tiene razón al afirmar que sus tartas se hornean en 5 minutos y no te importa si el tiempo medio real es mayor o menor que ese tiempo, utilizas la alternativa "distinto de". Tus hipótesis para este contraste serían H_0: $\mu=5$ contra H_a: $\mu \neq 5$.

Si únicamente quieres saber si el tiempo resulta ser mayor de lo que la empresa afirma (esto es si la empresa está mintiendo al decir en sus anuncios que sus tartas se preparan rápidamente), utilizas la alternativa "mayor que", en cuyo caso las dos hipótesis son H_0: $\mu=5$ contra H_a: $\mu>5$.

Por último, pongamos que trabajas para la empresa que comercializa esas tartas y crees que pueden hornearse incluso en menos de cinco minutos (y la empresa podría decirlo así en sus anuncios). En ese caso quieres la alternativa "menor que", y tus dos hipótesis serán H_0: $\mu=5$ contra H_a: $\mu<5$.

¿Cómo sabes qué hipótesis poner en H_0 y en H_a? La hipótesis nula dice que no ocurre nada nuevo; el resultado previo no varía o los grupos tienen la misma media (la diferencia es igual a cero). En general, se supone que una afirmación es verdadera hasta que se demuestre lo contrario. Así pues, la pregunta debe ser: ¿puedes demostrar lo contrario? Dicho de otro modo, ¿tienes pruebas suficientes para rechazar H_0?

Recopilar pruebas (datos)

Una vez formuladas las hipótesis, el siguiente paso consiste en recopilar pruebas y determinar si dichas pruebas son contrarias al enunciado de H_0. Recuerda que la afirmación se refiere a la población, pero es imposible contrastar a toda la población. Generalmente, lo mejor que puedes hacer es tomar una muestra. Como otras situaciones en que se recopilan estadísticas, la calidad de los datos es fundamental. (En el capítulo 3 explico cómo detectar estadísticas erróneas.)

Para recopilar datos válidos, lo primero es seleccionar una muestra válida. A la hora de seleccionar una muestra debes tener en cuenta dos cosas: evitar el sesgo y ser preciso. Para evitar el sesgo al seleccionar la muestra, asegúrate de que es aleatoria (los puntos de muestreo tienen la misma probabilidad de ser seleccionados que cualquiera otros posibles en la misma cantidad) y elige un tamaño muestral suficientemente grande como para que los resultados sean precisos. (En el capítulo 11 encontrarás más información sobre la precisión.)

Hay muchas formas distintas de recopilar datos, pero los métodos utilizados se reducen básicamente a dos: encuestas (estudios de observación) y experimentos (estudios controlados). En el capítulo 16 encontrarás todo lo que necesitas saber para diseñar y valorar encuestas, y para seleccionar muestras correctamente. En el capítulo 17 hablo de los experimentos: qué ventajas tienen sobre un estudio de observación, cuáles son los criterios de un experimento válido y cuándo puede inferirse una relación causa-efecto.

Procesar las pruebas: el estadístico de contraste

Una vez seleccionada la muestra, comienzan los cálculos. La hipótesis nula (H_0) plantea una afirmación referida al parámetro poblacional; por ejemplo "La proporción de mujeres estadounidenses que tienen varices es 0,25" (dicho de otro modo, H_0: p=0,25), o "El promedio de kilómetros que puede recorrer una furgoneta con un litro de gasolina es 9" (H_0: μ=9). Los datos recopilados con la muestra miden la variable de interés, y los estadísticos calculados te ayudan a contrastar la afirmación referida al parámetro poblacional.

Recopilar estadísticos muestrales

Supongamos que estás contrastando una afirmación sobre la proporción de mujeres que tienen varices. En ese caso tienes que calcular la proporción de mujeres de tu muestra que tienen varices, y ese número será tu estadístico muestral. Si estás contrastando una afirmación referida al promedio de kilómetros que puede recorrer una furgoneta con un litro de gasolina, tu estadístico será el promedio de kilómetros que recorran las furgonetas de tu muestra con un litro de gasolina. Y sabiendo que quieres medir la variabilidad en cuanto al promedio de kilómetros por litro para varias furgonetas, tienes que calcular la desviación estándar de la muestra. (En el capítulo 5 encontrarás toda la información que necesitas sobre el cálculo de estadísticos muestrales.)

Medir la variabilidad con errores estándares

Una vez que hayas calculado todos los estadísticos muestrales necesarios, quizá creas que has terminado con la parte de análisis y que ya estás listo para extraer conclusiones, pero no es así. El problema es que los resultados están en sus unidades originales y eso te impide ver las cosas con perspectiva. Sabes que los resultados se basan en una muestra, y seguro que esos resultados muestrales van a variar. Si no tienes en cuenta esa variación, las conclusiones podrían estar totalmente equivocadas. (¿Cuánto varían los resultados muestrales? La variación muestral se mide con el error estándar; encontrarás más información al respecto en el capítulo 11.)

Supongamos que la afirmación es que el porcentaje de mujeres que tienen varices es el 25%, y en tu muestra de 100 mujeres el 20% las padecía. El error estándar de tu porcentaje muestral es el 4% (según las fórmulas del capítulo 11), de manera que los resultados variarán aproximadamente el doble de ese valor, es decir, alrededor del 8%, según la regla empírica (está en el capítulo 12). Por tanto, una diferencia del 5%, por ejemplo, entre la afirmación y tu resultado muestral (25%–20%=5%) tampoco es mucho, porque representa una divergencia inferior a dos errores estándares respecto de la afirmación.

Sin embargo, si tu porcentaje muestral está basado en una muestra de 1.000 mujeres, en lugar de 100, la cosa cambia. En este caso, al tener más información, esperas que tus resultados varíen menos. Si aplicas de nuevo las fórmulas del capítulo 11, esta vez te sale un error estándar de 0,013, el 1,3%. El margen de error (ME) es aproximadamente el doble, es decir, un 2,6% por ambos lados. Ahora una diferencia del 5% entre tu resultado muestral (20%) y la afirmación representada por H_0 (25%) es una diferencia más significativa, ya que supera con mucho dos errores estándares.

¿Qué significación tienen tus resultados exactamente? En el siguiente apartado explico con más detalle la manera de medir con exactitud cuál es la divergencia entre tus resultados muestrales y la afirmación, en número de errores estándares. De este modo puedes extraer una conclusión específica sobre las pruebas que tienes en contra de la afirmación representada por H_0.

Comprender las puntuaciones estándares

El número de errores estándares que un estadístico está por encima o por debajo de la media se llama *puntuación estándar* (por ejemplo, un valor *z* es un tipo de puntuación estándar; se explica en el capítulo 9). A fin de interpretar tu estadístico, tienes que convertirlo de unidades originales a una puntuación estándar. Para obtener una puntuación estándar, coges el estadístico, le restas la media y divides el resultado por el error estándar.

En el caso de los contrastes de hipótesis, utilizas el valor de H_0 como media (te quedas con ese valor a menos que/hasta que tengas pruebas suficientes en su contra). La versión normalizada de tu estadístico se llama *estadístico de contraste*, y es el principal componente de un contraste de hipótesis. (En el capítulo 15 encontrarás las fórmulas para los contrastes de hipótesis más frecuentes.)

Calcular e interpretar el estadístico de contraste

El procedimiento general para convertir un estadístico en un estadístico de contraste (puntuación estándar) es el siguiente:

1. **Resta al estadístico el valor propuesto (el número mencionado en H_0).**

2. **Divide por el error estándar del estadístico.** (Existen diferentes fórmulas de error estándar para diferentes problemas; en el capítulo 13 encontrarás fórmulas detalladas para el error estándar, y en el capítulo 15, fórmulas para varios estadísticos de contraste.)

Tu estadístico de contraste representa la divergencia entre tus resultados muestrales y el valor poblacional propuesto, en número de errores estándares. En el caso de una única media o proporción poblacional, sabes que estas divergencias estandarizadas deberían tener una distribución al menos aproximadamente normal si el tamaño muestral es suficientemente grande (hablamos de ello en el capítulo 11). Así pues, para interpretar tu estadístico de contraste en estos casos puedes ver en qué lugar de la distribución normal estándar (distribución *Z*) se encuentra.

Si utilizamos los números del ejemplo anterior sobre mujeres con varices, para obtener el estadístico de contraste tienes que coger la proporción de mujeres de la muestra que tienen varices (0,2), restar la proporción propuesta de mujeres con varices (0,25), y luego dividir el resultado por el error estándar (0,04). Según estos cálculos, el estadístico de contraste (puntuación estándar) es –0,05/0,04 = –1,25. Esto te dice que tus resultados muestrales y la afirmación representada por H_0, referida a la población, presentan una divergencia de 1,25 errores estándares; concretamente, que tus resultados muestrales están 1,25 errores estándares por debajo del valor propuesto. Ahora debes preguntarte si esto es prueba suficiente para rechazar la afirmación. Encontrarás la respuesta a esa pregunta en el siguiente apartado.

Ponderar los indicios y tomar decisiones: valores *p*

Una vez has hallado el estadístico de contraste, lo utilizas para decidir si debes rechazar o no H_0. Para tomar esta decisión te basas en un número que mide la fuerza de este indicio (tu estadístico de contraste) contra la afirmación representada por H_0. Es decir, ¿qué probabilidad hay de que tu estadístico de contraste haya ocurrido siendo verdadera la afirmación? Este número que calculas se llama *valor p*; es la probabilidad de que alguien pudiera haber obtenido resultados tan extremos como los tuyos aunque H_0 sea correcta. Es la misma situación que se produce durante las deliberaciones tras un juicio: los miembros del jurado tienen que determinar qué probabilidad hay de que las pruebas sean las que son suponiendo que el acusado es inocente.

En este apartado hablo sobre los entresijos de los valores *p*, incluida la forma de calcularlos y utilizarlos para tomar decisiones en relación con H_0.

Relacionar estadísticos de contraste y valores *p*

Para comprobar si la afirmación representada por H_0 debe rechazarse (al fin y al cabo, todo gira en torno a H_0) tienes que considerar el estadístico de contraste que has obtenido de tu muestra y ver si tienes pruebas suficientes para rechazar la afirmación. Si el estadístico de contraste es grande (en sentido positivo o negativo), tus datos diferirán mucho de la afirmación (estarán muy alejados de ella); es decir, cuanto más grande sea el estadístico de contraste, más indicios tendrás en contra de la afirmación. Para saber cuándo puedes considerar que tus datos "están muy alejados", fíjate en dónde se encuentra tu estadístico de contraste dentro de la distribución de la cual procede. Cuando contrastas una media poblacional, si se cumplen ciertas condiciones, la distribución con la cual realizas la comparación es la distribución normal estándar (distribución *Z*), cuya media es 0 y cuya desviación estándar es 1; ésta es la distribución que utilizo como ejemplo a lo largo de este apartado. (En el capítulo 9 encontrarás más información sobre la distribución *Z*.)

Si el estadístico de contraste está cerca de cero, o al menos dentro de la zona donde deberían estar la mayoría de los resultados, entonces no tienes muchos indicios en contra de la afirmación (H_0), basándote en tus datos. Si tu estadístico de contraste se encuentra en las colas de la distribución normal estándar (más información en el capítulo 9), entonces hay muchos indicios en contra de la afirmación (H_0); este resultado tiene una probabilidad muy pequeña de ocurrir si la afirmación es verdadera. Dicho de otro modo; tienes suficientes indicios en contra de la afirmación (H_0) y, por tanto, la rechazas.

Pero ¿cuándo se considera que el valor está demasiado lejos del cero? Si tienes una distribución normal o un tamaño muestral suficientemente grande, sabes que tu estadístico de contraste estará en algún lugar de una distribución normal estándar (tratada en el capítulo 11). Si la hipótesis nula (H_0) es verdadera, la mayoría (más o menos el 95%) de las muestras generarán estadísticos de contraste que se encontrarán aproximadamente a no más de dos errores estándares de la afirmación. Si H_a es la alternativa "distinto de", cualquier estadístico de contraste que se encuentre fuera de este rango dará lugar a un rechazo de H_0. La figura 14-1 muestra las ubicaciones posibles del estadístico de contraste y las conclusiones correspondientes. En el siguiente apartado verás cómo se cuantifican los indicios que tienes en contra de H_0.

Observa que si la hipótesis alternativa es la alternativa "menor que", únicamente rechazarás H_0 si el estadístico de contraste se encuentra en la cola izquierda de la distribución (por debajo de –1,64). De forma similar, si H_a es la alternativa "mayor que", únicamente rechazarás H_0 si el estadístico de contraste se encuentra en la cola derecha (por encima de 1,64).

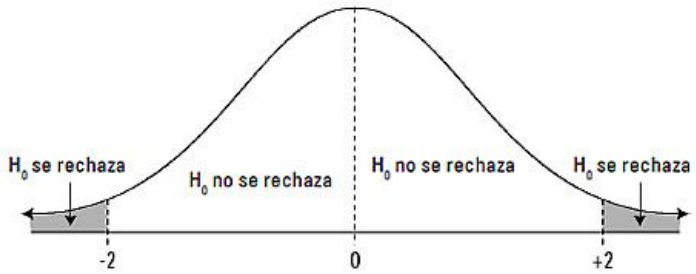


Figura 14-1: Decisiones para H_a: distinto de

Definir un valor *p*

Un valor *p* es una probabilidad asociada al estadístico de contraste. Mide la probabilidad de obtener resultados al menos tan extremos como el tuyo si la afirmación (H_0) es verdadera. Si estás contrastando la media poblacional, cuanto más hacia el extremo se encuentre tu estadístico de contraste en las colas de la distribución normal estándar (distribución *Z*), más pequeño será tu valor *p* y, por tanto, menos probable será que los resultados puedan haber ocurrido y más indicios tendrás en contra de la afirmación (H_0).

Calcular un valor *p*

Para hallar el valor *p* para un estadístico de contraste:

1. **Busca el estadístico de contraste en la distribución apropiada, en este caso la distribución normal estándar (distribución *Z*) (encontrarás la tabla *Z* en el apéndice).**

2. **Encuentra la probabilidad de que *Z* esté en una posición más extrema que tu estadístico de contraste:**

   - Si H_a contiene una alternativa "menor que" (contraste unilateral izquierdo), busca la probabilidad de que *Z* sea más pequeña que el estadístico de contraste (es decir, busca tu estadístico de contraste en la tabla *Z* y mira cuál es la probabilidad correspondiente). Ése es el valor *p*.

   - Si H_a contiene una alternativa "mayor que" (contraste unilateral derecho), busca la probabilidad de que *Z* sea más grande que el estadístico de contraste (es decir, busca tu estadístico de contraste en la tabla *Z*, mira cuál es la probabilidad que le corresponde y réstala de uno). El resultado es tu valor *p*.

   - Si H_a contiene una alternativa "distinto de" (contraste bilateral), busca la probabilidad de que *Z* no coincida con el estadístico de contraste y multiplícala por dos. Existen dos casos:

   - Si el estadístico de contraste es negativo, encuentra primero la probabilidad de que *Z* sea más pequeña que él (busca tu estadístico de contraste en la tabla *Z* y mira cuál es la probabilidad correspondiente). A continuación multiplica por dos esa probabilidad para obtener el valor *p*.

   - Si el estadístico de contraste es positivo, encuentra primero la probabilidad de que *Z* sea más grande que él (busca tu estadístico de contraste en la tabla *Z*, mira cuál es la probabilidad que le corresponde y réstala de uno). A continuación multiplica por dos ese resultado para obtener el valor *p*.

¿Por qué multiplicas por dos las probabilidades cuando H_a contiene una alternativa del tipo "distinto de"? La alternativa "distinto de" viene a ser la combinación de la alternativa "mayor que" y la alternativa "menor que". Si tienes un estadístico de contraste positivo, su valor *p* únicamente tiene en cuenta la parte "mayor que" de la alternativa "distinto de", de manera que hay que doblarlo para tener en cuenta también la parte "menor que" (es posible doblar el valor *p* porque la distribución *Z* es simétrica).

De forma similar, si tienes un estadístico de contraste negativo, su valor *p* únicamente tiene en cuenta la parte "menor que" de la alternativa "distinto de", y tienes que doblarlo para tener en cuenta también la parte "mayor que".

Volviendo al ejemplo anterior de las mujeres con varices, si contrastas H_0: $p = 0{,}25$ contra H_a: $p < 0{,}25$, el valor *p* resulta ser 0,1056. Esto es así porque el estadístico de contraste (calculado en el apartado anterior) era –1,25, y cuando buscas este número en la tabla *Z* (incluida en el apéndice) ves que la probabilidad de estar por debajo de este valor es 0,1056. Si hubieras contrastado la alternativa bilateral, H_a: $p \neq 0{,}25$, el valor *p* sería $2 \times 0{,}1056$, es decir, 0,2112.

Si es probable que los resultados hayan ocurrido siendo verdadera la afirmación, entonces no rechazas H_0 (igual que el jurado declara al acusado no culpable). Si es improbable que los resultados hayan ocurrido siendo verdadera la afirmación, entonces sí rechazas H_0 (igual que el jurado declara al acusado culpable). El punto de corte entre rechazar H_0 y no rechazar H_0 es un asunto peliagudo al que me refiero en el siguiente apartado.

Extraer conclusiones

Para extraer conclusiones sobre H_0 (rechazarla o no rechazarla) a partir de un valor *p*, tienes que definir un punto de corte predeterminado de manera que únicamente rechazarás H_0 si el valor *p* es menor o igual que ese punto de corte. Este punto de corte se llama *nivel alfa (α)* o *nivel de*

significación para el contraste. Si bien el valor 0,05 es un punto de corte habitual para rechazar H_0, los puntos de corte y las decisiones resultantes pueden variar: hay quien utiliza un punto de corte más estricto (por ejemplo 0,01), que requiere indicios más claros antes de rechazar H_0, y hay quien prefiere un punto de corte menos estricto (por ejemplo 0,10), que requiere menos indicios.

Si H_0 se rechaza (es decir, el valor *p* es menor o igual que el nivel de significación predeterminado), el investigador puede decir que ha encontrado un resultado estadísticamente significativo. Un resultado es *estadísticamente significativo* (o tiene *significación estadística*) cuando es demasiado poco común como para deberse al azar suponiendo que H_0 sea verdadera. Si obtienes un resultado estadísticamente significativo, tienes suficientes indicios como para rechazar la afirmación, H_0, y concluir que la realidad es otra (es decir, H_a).

- ✔ Si el valor *p* es menor o igual que el nivel de significación, entonces cumple tus requisitos para tener suficientes indicios contra H_0; por tanto, rechazas H_0.
- ✔ Si el valor *p* es mayor que el nivel de significación, los datos no aportan pruebas que estén más allá de toda duda razonable; por tanto, no rechazas H_0.

Sin embargo, si tienes previsto tomar decisiones sobre H_0 comparando el valor *p* con tu nivel de significación, tienes que definir previamente ese nivel de significación. No sería justo cambiar el punto de corte después de haber visto los datos y saber por dónde van los tiros.

A lo mejor te estás preguntando si puedes decir "aceptar H_0" en lugar de "no rechazar H_0". La respuesta es no. En un contraste de hipótesis, no intentas demostrar si H_0 es verdadera o no (que es lo que implica la palabra "aceptar"); de hecho, si supieras que H_0 es verdadera, ya no harías el contraste de hipótesis. De lo que se trata es de ver si tienes suficientes indicios para decir que H_0 es falsa, basándote en tus datos. Entonces, o tienes suficientes indicios para decir que es falsa (en cuyo caso rechazas H_0) o no tienes suficientes indicios para decir que es falsa (en cuyo caso no rechazas H_0).

Establecer límites para rechazar H_0

Estas directrices te ayudarán a tomar una decisión (rechazar o no rechazar H_0) basándote en un valor *p* cuando el nivel de significación sea 0,05:

- ✔ Si el valor *p* es menor que 0,01 (muy pequeño), se considera que los resultados tienen una elevada significación estadística. Por tanto, rechazas H_0.
- ✔ Si el valor *p* está entre 0,05 y 0,01 (pero no extremadamente cerca de 0,05), se considera que los resultados tienen significación estadística. Por tanto, rechazas H_0.
- ✔ Si el valor *p* está muy cerca de 0,05 (por ejemplo 0,051 o 0,049), debe considerarse que los resultados tienen cierta significación estadística. En ese caso puedes tomar cualquiera de las dos decisiones.
- ✔ Si el valor *p* es mayor que 0,05 (pero no está extremadamente cerca de 0,05), se considera que los resultados no tienen significación estadística. Por tanto, no rechazas H_0.

Si oyes que un investigador dice que sus resultados tienen significación estadística, busca el valor *p* y decide tú si la tienen o no. Puede que el nivel de significación predeterminado del investigador no coincida con el tuyo. Si no mencionan el valor *p*, pídelo.

Un problema de varices

En el ejemplo anterior sobre las mujeres con varices, vimos que el valor *p* era 0,1056. Este valor *p* es bastante grande y aporta indicios muy débiles en contra de H_0 para casi cualquier nivel de exigencia, ya que es mayor que 0,05 e incluso un poco mayor que 0,10 (que se considera un nivel de significación muy alto). En este caso no rechazas H_0. No había indicios suficientes para decir que la proporción de mujeres con varices es inferior a 0,25 (la hipótesis alternativa). No se considera un resultado estadísticamente significativo.

Pero ahora supongamos que tu valor *p* hubiera sido 0,026. Un lector que hubiera establecido su punto de corte personal en 0,05 rechazaría H_0 en este caso porque el valor *p* (0,026) es menor que 0,05. Su conclusión sería que la proporción de mujeres con varices no es igual a 0,25; según

H_a en este caso, concluiría que es menor que 0,25, y los resultados serían estadísticamente significativos. Sin embargo, un lector cuyo nivel de significación fuera 0,01 no tendría indicios suficientes (basándose en tu muestra) para rechazar H_0 porque el valor *p* de 0,026 es mayor que 0,01. Estos resultados no serían estadísticamente significativos. Por último, si el valor *p* fuera 0,049 y tu nivel de significación fuera 0,05, podrías seguir las reglas y decir que, por ser un valor menor que 0,05, rechazas H_0, pero tendrías la obligación de mencionar que tus resultados no son concluyentes, para que fuera el lector quien decidiera (quizá podría lanzar una moneda al aire: "Cara, rechazo H_0; cruz, no la rechazo").

Valorar la posibilidad de tomar una decisión errónea

Una vez que has tomado la decisión de rechazar H_0 o no rechazar H_0, el siguiente paso es afrontar las consecuencias, es decir, cómo responde la gente a tu decisión.

- ✔ Si concluyes que una afirmación no es cierta pero en realidad sí lo es, ¿el resultado será una demanda, cambios innecesarios en el producto o un boicot de los consumidores que no debería haberse producido? Es posible.
- ✔ Si no puedes refutar una afirmación falsa, ¿qué ocurre a continuación? ¿Continuarán fabricándose los productos igual que hasta ahora? ¿No se dictará una nueva ley ni se adoptarán nuevas medidas porque no has demostrado que haya que cambiar nada? No sería la primera vez que se ha perdido la oportunidad de pulsar el botón de alarma.

Decidas lo que decidas al realizar el contraste de hipótesis, sabes que puedes equivocarte. Las cosas son así en el mundo de la estadística. La clave está en saber qué tipos de errores pueden ocurrir y cómo reducir las probabilidades de que ocurran.

Provocar una falsa alarma: errores de tipo 1

Imagina que una empresa de mensajería afirma que el tiempo medio de entrega de sus paquetes es de dos días, y una organización de consumidores contrasta esta hipótesis, obtiene un valor *p* de 0,04 y concluye que esa afirmación es falsa: creen que el tiempo medio de entrega es en realidad superior a dos días. Esta situación es bastante seria. Si la organización de consumidores puede justificar su decisión, ha hecho bien en informar al público sobre este problema de publicidad engañosa. Pero ¿qué ocurre si la organización está equivocada?

Incluso si la organización diseña el estudio correctamente, recopila datos válidos y realiza bien el análisis, puede equivocarse. ¿Por qué? Porque sus conclusiones se basan en una muestra de paquetes, no en toda la población. Y en el capítulo 11 vimos que los resultados muestrales varían de una muestra a otra.

Que los resultados de una muestra se salgan de lo normal no significa que sean imposibles. Un valor *p* de 0,04 significa que la probabilidad de obtener tu estadístico de contraste en concreto, incluso si la afirmación es verdadera, es del 4% (inferior al 5%). En este caso rechazas H_0 porque esa probabilidad es pequeña. Pero ¡una probabilidad pequeña sigue siendo una probabilidad!

Puede que la muestra, a pesar de haberse seleccionado de forma aleatoria, sea una de esas muestras atípicas cuyo resultado termina alejándose mucho de lo que cabía esperar. Entonces, H_0 podría ser verdadera, pero tus resultados te han conducido a una conclusión diferente. ¿Con qué frecuencia ocurre esto? El 5% de las veces (o la probabilidad de corte que establezcas para rechazar H_0).

Rechazar H_0 siendo verdadera es lo que se denomina *error de tipo 1*. A mí este nombre no me gusta porque no es nada descriptivo. Prefiero llamarlo *falsa alarma*. En el caso de los paquetes, si la organización de consumidores cometió un error del tipo 1 al rechazar la afirmación de la empresa, provocó una falsa alarma. ¿Cuál es el resultado? Una empresa de mensajería muy enfadada, de eso puedes estar seguro.

Para reducir la probabilidad de crear falsas alarmas, define una probabilidad de corte (nivel de significación) baja para rechazar H_0. Si la fijas en el 5 o el 1%, la probabilidad de cometer un error de tipo 1 estará bajo control.

No detectar una afirmación falsa: errores de tipo 2

Ahora supongamos que la empresa realmente no entrega los paquetes en el plazo que dice. ¿Quién puede asegurar que la muestra de la organización de consumidores lo detectará? Si el tiempo de entrega real es de 2,1 días en lugar de 2, la diferencia sería muy difícil de detectar. Si el tiempo de entrega es de 3 días, incluso una muestra bastante pequeña probablemente podría descubrir que algo está pasando. El problema está en los valores intermedios, como por ejemplo 2,5 días.

Si H_0 es falsa, tienes que averiguarlo y rechazarla. No rechazar H_0 siendo falsa es lo que se denomina *error de tipo 2*. Yo prefiero llamarlo *error de detección*.

El tamaño muestral es la clave para detectar situaciones en que H_0 es falsa y, de este modo, evitar errores de tipo 2. Cuanta más información tengas, menos variables serán tus resultados (como se explica en el capítulo 11) y más probable será que consigas detectar algún problema en la afirmación representada por H_0.

Esta probabilidad de detectar cuándo H_0 es en realidad falsa se denomina la *potencia* de un contraste. La potencia es una cuestión bastante compleja, pero de momento te basta con saber que cuanto mayor es el tamaño muestral, más potente es el contraste. Un contraste potente conlleva una probabilidad baja de cometer un error de tipo 2.

Como medida de prevención para reducir al mínimo la probabilidad de cometer un error de tipo 2, los estadísticos recomiendan seleccionar un tamaño muestral grande para garantizar que las discrepancias respecto de la situación real no se pasarán por alto.

Capítulo 15

Contrastes de hipótesis frecuentemente utilizados: fórmulas y ejemplos

En este capítulo

- Analizar a fondo los contrastes de hipótesis más utilizados
- Calcular sus estadísticos de contraste
- Utilizar los resultados para tomar decisiones con conocimiento de causa

Desde los anuncios de productos hasta las noticias sobre recientes avances en medicina, nos encontramos a todas horas con afirmaciones referidas a una o más poblaciones. Por ejemplo: "Prometemos entregar su paquete en dos días o menos" o "Dos estudios recientes demuestran que una dieta rica en fibra puede reducir en el 20% el riesgo de padecer cáncer de colon". Cuando alguien hace una afirmación (también llamada *hipótesis nula*) sobre una población (por ejemplo todos los paquetes enviados, o todos los adultos), puedes contrastar esa afirmación mediante lo que en estadística se denomina un *contraste de hipótesis*.

Para realizar un contraste de hipótesis tienes que formular las *hipótesis* (una afirmación y su alternativa), recopilar datos, calcular los estadísticos relevantes y utilizar esos estadísticos para decidir si la afirmación es verdadera.

En este capítulo te presento las fórmulas utilizadas para los contrastes de hipótesis más comunes, explico los cálculos necesarios y comento algunos ejemplos.

Si necesitas más información sobre contrastes de hipótesis (definir hipótesis, comprender los estadísticos de contraste, valores *p*, niveles de significación y errores de tipo 1 y de tipo 2), ve al capítulo 14. Allí desarrollo todos los conceptos generales de los contrastes de hipótesis. Este capítulo se centra en su aplicación.

Contrastar una media poblacional

Cuando la variable es numérica (por ejemplo edad, ingresos, tiempo, etc.) y únicamente se estudia una población o grupo (por ejemplo todas las familias de Estados Unidos o todos los estudiantes universitarios), para analizar o cuestionar una afirmación referida a la media poblacional se utiliza el contraste de hipótesis descrito en este apartado. Por ejemplo, un psicólogo infantil dice que las madres trabajadoras dedican, en promedio, 11 minutos diarios a hablar con sus hijos (según ese mismo psicólogo, los padres dedican un tiempo medio de 8 minutos). La variable (el tiempo) es numérica, y la población son todas las madres trabajadoras. Si utilizamos la notación estadística, μ representa el número de minutos diarios que todas las madres trabajadoras dedican a hablar con sus hijos, en promedio.

La hipótesis nula es que la media poblacional, μ, es igual a un determinado valor propuesto, μ_0. La notación de la hipótesis nula es H_0: $\mu = \mu_0$. Así pues, la hipótesis nula de nuestro ejemplo es H_0: $\mu = 11$ minutos, y μ_0 es 11. Las tres posibilidades para la hipótesis alternativa, H_a, son $\mu \neq 11$, $\mu < 11$ y $\mu > 11$, dependiendo de lo que pretendas demostrar (en el capítulo 14 encontrarás más información sobre hipótesis alternativas). Si sospechas que el tiempo medio que dedican las madres a hablar con sus hijos es superior a 11 minutos, elegirás la hipótesis alternativa H_a: $\mu > 11$.

Para contrastar la afirmación, comparas la media obtenida a partir de tu muestra ($\bar{x}$) con la media indicada en H_0 (μ_0). Para que la comparación sea correcta, determinas la diferencia entre ambos valores y divides por el error estándar para tener en cuenta el hecho de que los resultados muestrales están sujetos a variación. (En el capítulo 12 encontrarás toda la información que necesitas sobre el error estándar.) Este resultado es tu *estadístico de contraste*. En el caso de un contraste de hipótesis para la media poblacional, el estadístico de contraste es (si se cumplen ciertas condiciones) un valor *z* (un valor de la distribución *Z*; ver el capítulo 9).

Por tanto, puedes buscar tu estadístico de contraste en la tabla apropiada (en este caso la tabla *Z*, incluida en el apéndice) y ver cuál es la probabilidad de que esta diferencia entre tu media muestral y la media poblacional propuesta pueda ocurrir realmente si la afirmación es verdadera.

El estadístico de contraste para contrastar una media poblacional (si se cumplen ciertas condiciones) es:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

donde $\bar{x}$ es la media muestral, σ es la desviación estándar de la población (en este caso supondremos que es un número conocido), y *z* es un valor de la distribución *Z*. Para calcular el estadístico de contraste, haz lo siguiente:

1. **Calcula la media muestral, $\bar{x}$.**

2. Encuentra $\bar{x}$ - μ_0.

3. Calcula el error estándar: $\sigma/\sqrt{n}$.

4. Divide el resultado del paso 2 por el error estándar calculado en el paso 3.

Las condiciones para utilizar este estadístico de contraste son, primero, que se conozca la desviación estándar de la población, σ, y segundo, que la población tenga una distribución normal o bien el tamaño muestral sea suficientemente grande para aplicar el teorema del límite central (n>30); sobre ello hablamos en el capítulo 11.

Para nuestro ejemplo, supondremos que una muestra aleatoria de 100 madres trabajadoras dedicaron una media de 11,5 minutos diarios a hablar con sus hijos (supondremos también que, según estudios anteriores, la desviación estándar de la población es 2,3 minutos).

1. Nos dicen que $\bar{x}$ es 11,5, n = 100 y σ es 2,3.
2. Calcula 11,5 – 11 = +0,5.
3. Para calcular el error estándar divides 2,3 por la raíz cuadrada de 100 (que es 10), y te da 0,23.
4. Ahora divides +0,5 por 0,23. El resultado de esta operación es 2,17. Éste es tu estadístico de contraste. Significa que tu media muestral está 2,17 errores estándares por encima de la media poblacional propuesta.

El contraste de hipótesis cuestiona la afirmación que se está haciendo sobre la población (en este caso, la media poblacional); esa afirmación está representada por la hipótesis nula, H_0. Si la muestra te proporciona suficientes indicios en contra de la afirmación, rechazas H_0.

Para decidir si tienes indicios suficientes para rechazar H_0, calcula el valor p buscando tu estadístico de contraste (en este caso 2,17) en la distribución normal estándar (distribución Z) —consulta la tabla Z en el apéndice— y resta de 1 la probabilidad hallada (restas de 1 porque tu H_a es una hipótesis del tipo "mayor que", y la tabla contiene probabilidades del tipo "menor que").

Para este ejemplo, si buscas el estadístico de contraste (2,17) en la tabla Z verás que la probabilidad ("menor que") es 0,985, de manera que el valor p es 1 –0,985 = 0,015. Este valor está bastante por debajo del nivel de significación (típico) de 0,05, lo cual significa que tus resultados muestrales se considerarían fuera de lo normal si la afirmación (de 11 minutos) fuera verdadera. Por tanto, rechazas la afirmación (H_0: μ = 11 minutos). Tus resultados apoyan la hipótesis alternativa H_a: μ > 11. Según los datos que has obtenido, los 11 minutos diarios que según el psicólogo infantil dedican las madres a hablar con sus hijos son un valor demasiado bajo; la media real es más alta.

En el capítulo 14 encontrarás más información sobre cómo calcular valores p para las alternativas "menor que" y "distinto de".

Muestras pequeñas y desviaciones estándar desconocidas: la prueba *t*

Hay dos casos en los que no puedes utilizar la distribución Z para un estadístico de contraste cuando quieres contrastar una media poblacional. El primer caso es cuando el tamaño muestral es pequeño (y cuando digo pequeño, quiero decir inferior a 30, aproximadamente), y el segundo caso es cuando la desviación estándar de la población, σ, se desconoce, y entonces hay que estimarla utilizando la desviación estándar de la muestra, s. En ambos casos tienes menos información fiable sobre la cual basar tus conclusiones, de manera que debes pagar un peaje utilizando una distribución que presenta más variabilidad en las colas que la distribución Z. Aquí es donde entra en escena la distribución t. (En el capítulo 10 encontrarás más información sobre la distribución t, incluida su relación con la distribución Z.)

Un contraste de hipótesis para una media poblacional que utilice la distribución t se llama prueba t. La fórmula para el estadístico de contraste en este caso es:

$$t_{n-1} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}},$$

en la que t_{n-1} es un valor de la distribución t que tiene n–1 grados de libertad.

Si te fijas, verás que es igual que el estadístico de contraste para el caso de una muestra grande o una distribución normal (lee el apartado "Contrastar una media poblacional"), salvo que σ es desconocida, de manera que la sustituyes por la desviación estándar de la muestra, s, y utilizas un valor t en lugar de un valor z.

Como la distribución t tiene las colas más gruesas que la distribución Z, el valor p obtenido con la distribución t es más grande que el que habrías obtenido con la distribución normal estándar (distribución Z) para el mismo estadístico de contraste. Un valor p más grande significa una

probabilidad menor de rechazar H_0. Tener pocos datos y no conocer la desviación estándar de la población aumenta la dificultad de determinar la hipótesis nula.

Aplicación práctica de la prueba *t*

Supongamos que una empresa de mensajería asegura que entrega sus paquetes en el plazo de dos días, en promedio, y tú sospechas que tarda más tiempo. Las hipótesis son H_0: $\mu = 2$ contra H_a: $\mu > 2$. Para contrastar esta afirmación, tomas una muestra aleatoria de 10 paquetes y anotas los tiempos de entrega. Averiguas que la media muestral es $\bar{x} = 2{,}3$ días y la desviación estándar de la muestra es 0,35 días (como desconoces la desviación estándar de la población, σ, la estimas con *s*, la desviación estándar de la muestra). Este problema se resuelve con la prueba *t*.

Como el tamaño muestral es pequeño (*n* =10 es mucho menos que 30) y la desviación estándar de la población es desconocida, el estadístico de contraste tiene una distribución *t*. Sus grados de libertad son 10 – 1 = 9. La fórmula para el estadístico de contraste (nos referiremos a él como *valor t*) es:

$$t_{10-1} = \frac{2{,}3 - 2{,}0}{0{,}35/\sqrt{10}} = 2{,}71$$

Para calcular el valor *p*, ve a la tabla *t* (incluida en el apéndice) y busca la fila que corresponde a $gl = 9$. Tu estadístico de contraste (2,71) se encuentra entre dos valores de la fila correspondiente a $gl = 9$ en la tabla *t*: 2,26 y 2,82 (redondeando a dos cifras decimales). Para calcular el valor *p* para tu estadístico de contraste, busca las columnas correspondientes a esos dos números. El número 2,26 aparece en la columna 0,025 y el número 2,82 aparece en la columna 0,010; por tanto, sabes que el valor *p* para tu estadístico de contraste se encuentra entre 0,025 y 0,010 (es decir, 0,010 < valor *p* < 0,025).

Utilizando la tabla *t* no puedes saber el número exacto que corresponde al valor *p*, pero como 0,01 y 0,025 son más pequeños que tu nivel de significación (0,05), rechazas H_0; la muestra te proporciona suficientes indicios para decir que los paquetes no se están entregando en dos días, sino que el tiempo medio de entrega es en realidad superior a dos días.

La tabla *t* (incluida en el apéndice) no incluye todos los valores *t* posibles; sólo tienes que buscar los dos números más próximos al tuyo por ambos lados y ver en qué columnas aparecen. El valor *p* se encuentra entre los valores *p* de esos números (si tu estadístico de contraste es más grande que todos los valores *t* de la fila correspondiente de la tabla *t*, utiliza el último; tu valor *p* será más pequeño que la probabilidad de ese último valor).

Naturalmente, puedes utilizar un programa informático, si lo tienes, para calcular los valores *p* exactos para cualquier estadístico de contraste; de este modo verías que el valor *p* exacto en este caso es 0,012.

Relacionar *t* y *Z*

La penúltima línea de la tabla *t* muestra los valores de la distribución normal estándar (distribución *Z*) que corresponden a las probabilidades indicadas en la cabecera de cada columna. Ahora elige una columna de la tabla y mira los valores *t*. Observa que, a medida que los grados de libertad de la distribución *t* aumentan, los valores *t* se acercan más y más a esa fila de la tabla donde están los valores *z*.

Esto confirma un resultado hallado en el capítulo 10: a medida que el tamaño muestral (y, por tanto, los grados de libertad) aumenta, la distribución *t* se asemeja cada vez más a la distribución *Z*, de modo que los valores *p* de sus contrastes de hipótesis son prácticamente iguales cuando el tamaño muestral es grande. Y ni siquiera es necesario que esos tamaños muestrales sean muy grandes para observar esta relación: para gl=30, los valores *t* ya son muy parecidos a los valores *z* indicados en la parte de abajo de la tabla. Estos resultados tienen todo el sentido: cuantos más datos tienes, menor es el peaje que tienes que pagar (y, por supuesto, puedes utilizar una aplicación informática para calcular valores *p* más exactos para cualquier valor *t* que desees).

Qué hacer con valores *t* negativos

En el caso de la hipótesis alternativa "menor que" (H_a: xx < xx), tu estadístico de contraste sería un número negativo (situado a la izquierda del cero en la distribución *t*). En este caso, para obtener el valor *p* tienes que encontrar el porcentaje que queda por debajo, o a la izquierda, de tu estadístico de contraste. Sin embargo, la tabla *t* (incluida en el apéndice) no contiene estadísticos de contraste negativos.

Tranquilo, no hay de qué preocuparse. Debido a la simetría, el porcentaje que hay a la izquierda (debajo) de un valor *t* negativo es igual que el

porcentaje que hay a la derecha (por encima) del valor *t* positivo. Por consiguiente, para determinar el valor *p* correspondiente a tu estadístico de contraste negativo, busca la versión positiva de tu estadístico de contraste en la tabla *t*, encuentra la probabilidad de cola derecha ("mayor que") correspondiente y utiliza ese valor.

Por ejemplo, supongamos que el estadístico de contraste es –2,7105 con 9 grados de libertad y H_a es la alternativa "menor que". Para determinar el valor *p*, primero buscas +2,7105 en la tabla *t*; según lo que hemos visto en el apartado anterior, sabes que su valor *p* está entre las cabeceras de columna 0,025 y 0,010. Como la distribución *t* es simétrica, el valor *p* para –2,7105 también está entre 0,025 y 0,010. De nuevo rechazas H_0 porque estos dos valores son menores o iguales que 0,05.

Analizar la alternativa "distinto de"

Para determinar el valor *p* cuando la hipótesis alternativa (H_a) es del tipo "distinto de", sólo tienes que doblar (multiplicar por dos) la probabilidad que obtienes de la tabla *t* cuando buscas tu estadístico de contraste. ¿Por qué razón la doblas? Porque la tabla *t* únicamente muestra probabilidades del tipo "mayor que", que sólo te cuentan la mitad de la película. Para encontrar el valor *p* cuando tienes una alternativa "distinto de", debes sumar los valores *p* de las alternativas "menor que" y "mayor que". Como la distribución *t* es simétrica, las probabilidades "menor que" y "mayor que" son idénticas, de manera que doblas la que has encontrado en la tabla *t* y ya tienes el valor *p* para la alternativa "distinto de".

Por ejemplo, si el estadístico de contraste es 2,7171 y H_a es una alternativa del tipo "distinto de", buscas 2,7171 en la tabla *t* (*gl*=9 otra vez) y compruebas que el valor *p* está entre 0,025 y 0,010, como hemos visto anteriormente. Éstos son los valores *p* para la alternativa "mayor que". Ahora doblas esos valores para incluir la alternativa "menor que" y obtienes que el valor *p* para tu estadístico de contraste está entre 0,025 × 2=0,05 y 0,01 × 2=0,02.

Contrastar una proporción poblacional

Cuando la variable es categórica (por ejemplo el sexo de la persona, o si está a favor/en contra de una determinada cuestión) y únicamente se estudia una población o grupo (por ejemplo todos los votantes censados), para contrastar una afirmación referida a la proporción poblacional se utiliza el contraste de hipótesis descrito en este apartado. El contraste examina la proporción *(p)* de elementos de la población que poseen cierta característica, por ejemplo la proporción de personas que tienen teléfono móvil. La hipótesis nula es H_0: $p = p_0$, donde p_0 es cierto valor propuesto para la proporción poblacional, *p*. Por ejemplo, si la afirmación es que el 70% de las personas tienen teléfono móvil, p_0 es 0,70. La hipótesis alternativa es una de las siguientes: $p > p_0$, $p < p_0$, o $p \neq p_0$. (En el capítulo 14 encontrarás más información sobre las hipótesis alternativas.)

La fórmula del estadístico de contraste para una única proporción (si se cumplen ciertas condiciones) es:

$$z = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}}$$

en la $\hat{p}$ que es la proporción de elementos de la muestra que tienen esa característica, y *z* es un valor de la distribución *Z* (ver el capítulo 9). Para calcular el estadístico de contraste, haz lo siguiente:

1. **Calcula la proporción muestral, $\hat{p}$, para lo cual tienes que coger el número de personas de la muestra que poseen la característica de interés (por ejemplo, el número de personas de la muestra que tienen teléfono móvil) y dividirlo por *n*, el tamaño muestral.**
2. **Calcula $\hat{p}$ - p_0, donde p_0 es el valor indicado en H_0.**
3. **Calcula el error estándar, $\sqrt{\frac{p_o(1-p_o)}{n}}$.**
4. **Divide el resultado del paso 2 por el resultado del paso 3.**

Para interpretar tu estadístico de contraste, búscalo en la distribución normal estándar (distribución *Z*, incluida en el apéndice) y calcula el valor *p*. (En el capítulo 14 encontrarás más información sobre cómo calcular el valor *p*.)

Las condiciones para utilizar este estadístico de contraste son que $np_0 \geq 10$ y $n(1 - p_0) \geq 10$. (Hay más información en el capítulo 9.)

Por ejemplo, imagina que la marca Sincaries afirma que cuatro de cada cinco dentistas recomiendan la pasta de dientes Sincaries a sus pacientes. En este caso la población son todos los dentistas, y *p* es la proporción de ellos que recomiendan Sincaries. La afirmación es que *p* es igual a "cuatro de cada cinco", es decir, p_0 es 4/5=0,8. Tú sospechas que la proporción es en realidad inferior a 0,8. Tus hipótesis son H_0 : p=0,8 contra

H_a : p<0,8.

Supongamos que tomas una muestra de 200 pacientes odontológicos y 151 de ellos manifiestan que su dentista les ha recomendado la marca Sincaries. Para encontrar el estadístico de contraste para estos resultados, sigue los pasos siguientes:

1. Empieza con $\hat{p}=\frac{151}{200}$ = 0,755 y n = 200.
2. Como p_0 = 0,8, restas 0,755 – 0,8 = –0,045 (el numerador del estadístico de contraste).
3. El error estándar es $\sqrt{\frac{0,80(1-0,80)}{200}}$ = 0,028 (el denominador del estadístico de contraste).
4. El estadístico de contraste es $\frac{-0,045}{0,028}$ = –1,61.

Como el estadístico de contraste es negativo, significa que tus resultados muestrales están –1,61 errores estándares por debajo (menos que) del valor propuesto para la población. ¿Con qué frecuencia cabe esperar que se obtengan estos resultados si H_0 es verdadera? La probabilidad de alcanzar o superar –1,61 (en este caso, en sentido negativo) es 0,0537 (conserva el signo negativo y busca –1,61 en la tabla *Z*). Este resultado es tu valor *p* porque H_a es una hipótesis del tipo "menor que". (En el capítulo 14 encontrarás más información al respecto.)

Como el valor *p* es mayor que 0,05 (aunque sea por poco), no tienes indicios suficientes para rechazar H_0. Concluyes que la afirmación de que el 80% de los dentistas recomiendan Sincaries no puede ser rechazada, de acuerdo con tus datos. No obstante, es importante comunicar también el valor *p* para que los demás puedan tomar sus propias decisiones.

La letra *p* se utiliza con dos significados distintos en este capítulo: valor *p* y *p*. La letra *p* por sí sola se refiere a la proporción poblacional, no al valor *p*. No te confundas. Cuando comuniques un valor *p*, asegúrate de poner delante la palabra "valor", para que no se confunda con *p*, la proporción poblacional.

Comparar dos medias poblacionales (independientes)

Cuando la variable es numérica (por ejemplo ingresos, nivel de colesterol o kilómetros por litro de combustible) y se comparan dos poblaciones o grupos (por ejemplo hombres frente a mujeres), para contrastar una afirmación sobre la diferencia entre las medias de dichas poblaciones o grupos se utilizan los pasos descritos en este apartado (por ejemplo, ¿la diferencia entre las medias poblacionales es igual a cero, lo cual significa que las medias son idénticas?). A fin de recopilar los datos necesarios para este contraste es preciso seleccionar dos muestras aleatorias independientes (totalmente separadas), una de cada población.

La hipótesis nula es que las dos medias poblacionales son idénticas, o dicho de otro modo, que su diferencia es igual a cero. La notación de la hipótesis nula es H_0: $\mu_1 = \mu_2$, en la que μ_1 representa la media de la primera población y μ_2 representa la media de la segunda población.

También puedes expresar la hipótesis nula como H_0: $\mu_1 - \mu_2 = 0$, haciendo hincapié en que la diferencia es igual a cero si las medias son iguales.

La fórmula del estadístico de contraste para comparar dos medias (si se cumplen ciertas condiciones) es:

$$z=\frac{(\bar{x}_1-\bar{x}_2)-0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

Para calcularla, haz lo siguiente:

1. **Calcula las medias muestrales y (supondremos que nos dan las desviaciones estándares de las poblaciones, σ_1 y σ_2). n_1 y n_2 representan los dos tamaños muestrales (no es necesario que sean iguales).**

   Estos cálculos se explican en el capítulo 5.

2. **Calcula la diferencia entre las dos medias muestrales: $\bar{x}_1 - \bar{x}_2$.**

   Como la diferencia $\mu_1 - \mu_2$ es igual a cero si H_0 es verdadera, no es necesario incluirla en el numerador del estadístico de contraste. Sin embargo, si el valor de la diferencia es cualquier número distinto de cero, hay que restar ese valor en el numerador del estadístico de contraste.

3. **Calcula el error estándar por medio de la siguiente ecuación:**

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

4. **Divide el resultado del paso 2 por el resultado del paso 3.**

Para interpretar el estadístico de contraste, añade los dos pasos siguientes:

5. **Busca tu estadístico de contraste en la distribución normal estándar (distribución *Z*) (encontrarás la tabla *Z* en el apéndice) y calcula el valor *p*.**

   (En el capítulo 14 encontrarás más información sobre cómo calcular el valor *p*.)

6. **Compara el valor *p* con el nivel de significación, por ejemplo 0,05. Si es menor o igual que 0,05, rechazas H_0. De lo contrario no rechazas H_0.**

   (En el capítulo 14 encontrarás más información sobre los niveles de significación.)

Las condiciones para utilizar este contraste son, primero, que se conozcan las desviaciones estándares de las poblaciones, y segundo, que las dos poblaciones tengan una distribución normal o bien los dos tamaños muestrales sean suficientemente grandes para aplicar el teorema del límite central (lo tienes en el capítulo 11).

Por ejemplo, imagina que quieres comparar la absorción de dos marcas de papel de cocina (las llamaremos marca A y marca B). Puedes realizar esta comparación midiendo el número medio de gramos que puede absorber una hoja de cada marca antes de saturarse del todo. H_0 dice que la diferencia entre los promedios de absorción es cero (inexistente), y H_a dice que la diferencia no es cero. Dicho de otro modo, una marca es más absorbente que la otra. En notación estadística tenemos $H_0 = \mu_1 - \mu_2 = 0$ contra $H_a = \mu_1 - \mu_2 \neq 0$. Aquí no se indica qué papel de cocina es más absorbente, de manera que hay que utilizar la alternativa "distinto de" (vuelve al capítulo 14 si lo necesitas).

Supongamos que seleccionas una muestra aleatoria de 50 hojas de cada marca y mides la absorción de cada una de ellas. Supongamos que la absorción media de la marca A *(x_1)* en tu muestra es 3 gramos, y supongamos que la desviación estándar de la población es 0,9 gramos. En el caso de la marca B *(x_2)*, la absorción media es 3,5 gramos según los datos de tu muestra, y la desviación estándar de la población es 1,2 gramos. Realiza este contraste de hipótesis siguiendo los seis pasos anteriores:

1. Por la información que tienes, sabes que $\bar{x}_1 = 30$, $\sigma_1 = 9$, $\bar{x}_2 = 35$, $\sigma_2 = 12$, $n_1 = 50$ y $n_2 = 50$.

2. La diferencia entre las medias muestrales (marcas A–marca B) es = (30 − 35) = −5 gramos. (Una diferencia negativa simplemente significa que la media de la segunda muestra era más grande que la media de la primera.)

3. El error estándar es $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{9^2}{50} + \frac{12^2}{50}} = \sqrt{\frac{81}{50} + \frac{144}{50}} = 2{,}1213$.

4. Al dividir la diferencia, –5, por el error estándar, 2,1213, te da –2,36. Éste es tu estadístico de contraste.

5. Para encontrar el valor *p*, busca –2,36 en la distribución normal estándar (distribución *Z*). Encontrarás la tabla *Z* en el apéndice. La probabilidad de superar –2,36, en este caso en sentido negativo, es igual a 0,0091. Como H_a es una alternativa del tipo "distinto de", doblas este porcentaje, con lo que obtienes 2 × 0,0091 = 0,0182, tu valor *p*. (En el capítulo 14 encontrarás más información sobre la alternativa "distinto de".)

6. Este valor *p* es bastante más pequeño que 0,05. Por consiguiente, tienes indicios suficientes para rechazar H_0.

Tu conclusión es que, por la información de tus muestras, existe una diferencia estadísticamente significativa entre la capacidad de absorción de estas dos marcas de papel de cocina. La marca B es más absorbente porque tiene una media más alta (que la media de la marca A menos la media de la marca B sea un número negativo significa que la marca B tenía un valor más alto).

Si una de las muestras, o las dos, tienen un tamaño inferior a 30, utilizas la distribución *t* (el valor correspondiente a los grados de libertad es $n_1 - 1$ o $n_2 - 1$, lo que sea más pequeño) para encontrar el valor *p*. Si las desviaciones estándares de las poblaciones, σ_1 y σ_2, se desconocen, utilizas en su lugar las desviaciones estándares de las muestras, s_1 y s_2, y utilizas la distribución *t* que tiene los grados de libertad arriba mencionados. (En el capítulo 10 encontrarás más información sobre la distribución *t*.)

Contraste de hipótesis para una diferencia media (prueba *t* para datos apareados)

Puedes realizar un contraste de hipótesis para una diferencia media utilizando la prueba descrita en este apartado cuando la variable es numérica (por ejemplo ingresos, nivel de colesterol o kilómetros por litro de combustible) y los elementos de la muestra están apareados de algún modo en relación con variables relevantes como por ejemplo la edad o el peso, o bien los mismos elementos se utilizan dos veces (por ejemplo, en una prueba preliminar y una prueba posterior). Las pruebas de datos apareados generalmente se utilizan para estudios en que alguien quiere saber si un nuevo tratamiento, técnica o método funciona mejor que otro método existente, sin tener que preocuparse de otros factores relacionados con los sujetos que puedan influir en los resultados. (Más información en el capítulo 17.)

La diferencia media (lo que se verifica en este apartado) no es lo mismo que la diferencia entre las medias (lo que se verificaba en el apartado anterior).

- ✔ En el caso de la diferencia entre las medias, comparamos la diferencia existente entre las medias de dos muestras separadas con el fin de determinar cuál es la diferencia existente entre las medias de dos poblaciones distintas.
- ✔ En el caso de la diferencia media, emparejamos los sujetos de manera que podamos considerar que proceden de una única población, y el conjunto de diferencias medidas para cada sujeto (por ejemplo, prueba preliminar contra prueba posterior) se considera una sola muestra. Entonces el contraste de hipótesis se reduce a una prueba para una media poblacional (como ya he explicado anteriormente en este mismo capítulo).

Por ejemplo, imagina que un investigador quiere averiguar si enseñar a los alumnos a leer utilizando un juego de ordenador da mejores resultados que enseñarles con un método fonético de eficacia probada. El investigador selecciona 20 alumnos y los organiza en 10 parejas en función de su dominio de la lectura, edad, coeficiente intelectual, etc. A continuación selecciona aleatoriamente a un alumno de cada pareja y le enseña a leer con el método informático (lo llamaremos MI), mientras que el otro aprende con el método fonético (lo llamaremos MF). Al final del estudio, los dos alumnos hacen el mismo examen de lectura. Los datos figuran en la tabla 15-1.

Tabla 15-1. Puntuaciones para el método informático y para el método fonético

Pareja de alumnos	*Método informático*	*Método fonético*	*Diferencia (MI-MF)*
1	85	80	+5
2	80	80	0
3	95	88	+7
4	87	90	−3
5	78	72	+6
6	82	79	+3
7	57	50	+7
8	69	73	−4
9	73	78	−5
10	99	95	+4

Los datos originales están en parejas, pero a ti únicamente te interesa la diferencia entre las puntuaciones obtenidas por los dos integrantes de cada pareja (la puntuación del alumno que ha aprendido con el método informático menos la puntuación del alumno que ha aprendido con el método fonético), no las puntuaciones en sí mismas. Por tanto, tu nuevo conjunto de datos son las *diferencias apareadas* (las diferencias entre las parejas de puntuaciones). Sus valores se indican en la última columna de la tabla 15-1.

Al examinar las diferencias entre las parejas de observaciones, en realidad tienes un solo conjunto de datos y un solo contraste de hipótesis para una media poblacional. En este caso la hipótesis nula es que la media (de las diferencias apareadas) es 0, y la hipótesis alternativa es que la media (de las diferencias apareadas) es>0. Si los dos métodos de aprendizaje son iguales, la media de las diferencias apareadas debe ser 0. Si el método informático es mejor, la media de las diferencias apareadas debe ser positiva (la puntuación de los alumnos que han aprendido con el método informático es mayor que la de los alumnos que han aprendido con el método fonético).

La notación de la hipótesis nula es H_0: $\sigma_d = 0$, donde σ_d es la media de las diferencias apareadas para la población (la *d* del subíndice

simplemente te recuerda que estás trabajando con las diferencias apareadas).

La fórmula del estadístico de contraste para diferencias apareadas es $t_{n-1} = \frac{\bar{d} - 0}{s_d / \sqrt{n_d}}$, en la que $\bar{d}$ es la media de todas las diferencias apareadas halladas en la muestra, y t_{n-1} es un valor de la distribución *t* que tiene $n_d - 1$ grados de libertad (puedes releer el capítulo 10).

Utilizas una distribución *t* porque en la mayoría de los experimentos con datos apareados el tamaño de muestra es pequeño o la desviación estándar de la población σ_d se desconoce, de manera que se estima con s_d. (En el capítulo 10 encontrarás más información sobre la distribución *t*.)

Para calcular el estadístico de contraste para diferencias apareadas, haz lo siguiente:

1. **Para cada par de datos, coge el primer valor del par y réstale el segundo valor del par para encontrar la diferencia apareada.** Considera que las diferencias son tu nuevo conjunto de datos.

2. **Calcula la media, $\bar{d}$, y la desviación estándar, s_d, de todas las diferencias.**

3. **Siendo n_d el número de diferencias apareadas que tienes, calcula el error estándar:**

   $$s_d / \sqrt{n_d}$$

4. **Divide $\bar{d}$ por el error estándar calculado en el paso 3.**

Como σ_d es igual a cero si H_0 es verdadera, no es necesario incluirla en la fórmula para hallar el estadístico de contraste. Por eso en ocasiones el estadístico de contraste se expresa de la manera siguiente:

$$\frac{\bar{d} - 0}{s_d / \sqrt{n_d}} = \frac{\bar{d}}{s_d / \sqrt{n_d}}$$

En el ejemplo de las puntuaciones obtenidas en los exámenes de lectura, puedes utilizar los pasos anteriores para ver si el método informático es mejor para enseñar a leer a los alumnos.

Para encontrar el estadístico, sigue los pasos siguientes:

1. **Calcula las diferencias para cada par (se indican en la columna 4 de la tabla 15-1).**

   El signo de cada diferencia es importante: indica qué método funcionó mejor para ese par en concreto.

2. **Calcula la media y la desviación estándar de las diferencias calculadas en el paso 1.**

   Según mis cálculos, la media de las diferencias, $\bar{d}$, es igual a 2, y la desviación estándar, s_d, es igual a 4,64. Observa que aquí $n_d = 10$.

3. **El error estándar es $\frac{4,64}{\sqrt{10}} = 1,47$.**

   (Recuerda que aquí n_d es el número de pares, que es 10.)

4. **Si divides la media de las diferencias (paso 2) por el error estándar (paso 3) te sale 1,36, que es el estadístico de contraste.**

¿El resultado del paso 4 es suficiente para decir que la diferencia entre puntuaciones hallada en este experimento se aplica a toda la población? Como la desviación estándar de la población, σ, se desconoce y la has estimado con la desviación estándar de la muestra *(s)*, tienes que utilizar la distribución *t* en lugar de la distribución *Z* para encontrar el valor *p*. (Revisa el apartado "Muestras pequeñas y desviaciones estándares desconocidas: la prueba *t*", en este mismo capítulo.) Para calcular el valor *p*, ve a la tabla *t* (incluida en el apéndice) y busca 1,36 en la distribución *t* que tiene 10–1=9 grados de libertad.

El valor *p* en este caso es mayor que 0,05 porque 1,36 es menor que el valor 1,38 de la tabla (es decir, se encuentra a la izquierda de dicho valor), y por tanto su valor *p* es mayor que 0,1 (el valor *p* para la cabecera de columna correspondiente a 1,38).

Como el valor *p* es mayor que 0,05, no rechazas H_0; no tienes indicios suficientes de que la diferencia media de puntuaciones entre el método

informático y el método fonético sea significativamente mayor que cero.

Eso no significa necesariamente que no exista una diferencia real en la población formada por todos los alumnos. Sin embargo, el investigador no puede afirmar que el juego de ordenador es un método mejor basándose en esta muestra de 10 alumnos. (En el capítulo 14 encontrarás más información sobre la potencia de un contraste de hipótesis y su relación con el tamaño muestral.)

En muchos experimentos apareados, los conjuntos de datos son pequeños para reducir el coste económico del muestreo y el tiempo que requiere la realización de este tipo de estudios. Esto significa que generalmente se utiliza la distribución *t* (tienes la tabla *t* en el apéndice) en lugar de la distribución normal estándar (distribución *Z*, ver la tabla *Z* en el apéndice) para determinar el valor *p*.

Comparar dos proporciones poblacionales

Este contraste se utiliza cuando la variable es categórica (por ejemplo fumador/no fumador, demócrata/republicano, a favor/en contra de cierta opinión, etc.) y estás interesado en la proporción de personas que poseen una determinada característica, por ejemplo, la proporción de fumadores. En este caso se comparan dos poblaciones o grupos (por ejemplo, la proporción de mujeres fumadoras contra la de hombres fumadores).

Para realizar este contraste hay que seleccionar dos muestras aleatorias independientes (separadas), una de cada población. La hipótesis nula es que las dos proporciones poblacionales son idénticas, o dicho de otro modo, que su diferencia es igual a cero. La notación de la hipótesis nula es H_0: $p_1 = p_2$, donde p_1 es la proporción de la primera población y p_2 es la proporción de la segunda población.

Afirmar en H_0 que las dos proporciones son idénticas es lo mismo que decir que su diferencia es cero. Si empiezas con la ecuación $p_1 = p_2$ y restas p_2 en cada lado, te queda $p_1 - p_2 = 0$. Por lo tanto, puedes expresar la hipótesis nula de ambas formas.

La fórmula del estadístico de contraste para comparar dos proporciones (si se cumplen ciertas condiciones) es:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

en la que $\hat{p}_1$ es la proporción de elementos de la primera muestra que poseen la característica de interés, $\hat{p}_2$ es la proporción de elementos de la segunda muestra que poseen la característica de interés, $\hat{p}$ es la proporción de elementos de la muestra combinada (todos los elementos de la primera y la segunda muestra) que poseen la característica de interés, y *z* es un valor de la distribución *Z* (hablamos de ella en el capítulo 9). Para calcular el estadístico de contraste, haz lo siguiente:

1. **Calcula las proporciones muestrales $\hat{p}_1$ y $\hat{p}_2$ para cada muestra. n_1 y n_2 representan los dos tamaños muestrales (no es necesario que sean iguales).**

2. **Calcula la diferencia entre las dos proporciones muestrales, $\hat{p}_1 - \hat{p}_2$.**

3. **Calcula la proporción muestral $\hat{p}$ global , el número total de elementos de ambas muestras que poseen la característica de interés (por ejemplo, el número total de fumadores, ya sean hombres o mujeres, que hay en la muestra) dividido por el número total de elementos de ambas muestras ($n_1 + n_2$).**

4. **Calcula el error estándar:**

$$\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

5. **Divide el resultado del paso 2 por el resultado del paso 4. La respuesta es tu estadístico de contraste.**

Para interpretar el estadístico de contraste, búscalo en la distribución normal estándar (la tabla *Z* del apéndice), calcula el valor *p* y toma las decisiones que corresponda. (En el capítulo 14 encontrarás más información sobre valores *p*.)

Piensa ahora en los anuncios de medicamentos que salen en las revistas. La foto principal muestra un sol esplendoroso, campos de flores, gente sonriendo... todo el mundo es más feliz gracias a ese medicamento. La compañía farmacéutica afirma que sus medicamentos pueden reducir los síntomas de la alergia, ayudar a la gente a dormir mejor, bajar la tensión arterial o solucionar cualquier otro problema de salud. Quizá todo esto te parezca demasiado bueno para ser cierto, pero cuando pasas la página para ver el reverso del anuncio, te encuentras con la letra pequeña donde el laboratorio farmacéutico justifica sus afirmaciones (¡ahí es donde suelen estar las estadísticas!). En algún lugar de esa letra pequeña es probable que encuentres una tabla donde se indiquen los efectos adversos del medicamento en comparación con un *grupo de control* (sujetos que tomaron un medicamento ficticio), para establecer un parangón justo con los que realmente tomaron el medicamento (el *grupo experimental*; más información en el capítulo 17).

Si nos fijamos en la información de una combinación de anfetamina y dextroanfetamina indicada para el trastorno por déficit de atención e hiperactividad (TDAH), dice que 26 de los 374 sujetos (el 7%) a quienes se les administró el fármaco experimentaron vómitos como efecto secundario, frente a 8 de los 210 sujetos (el 4%) que tomaron un placebo. Ten presente que los pacientes no sabían qué tratamiento estaban recibiendo. De todas las personas de la muestra que tuvieron vómitos, los que tomaron el fármaco representaron un porcentaje más alto. Sin embargo, ¿es suficiente este porcentaje para afirmar que toda la población que lo tome tendrá vómitos? Puedes hacer un contraste de hipótesis para averiguarlo.

En este ejemplo tienes H_0: $p_1 - p_2 = 0$ contra H_a: $p_1 - p_2 > 0$, donde p_1 representa la proporción de sujetos que tuvieron vómitos al tomar el medicamento, y p_2 representa la proporción de sujetos que tuvieron vómitos tomando el placebo.

¿Por qué H_a contiene el signo > y no el signo <? H_a representa la situación en la que los sujetos que tomaron el fármaco tuvieron más vómitos que los que tomaron el placebo, algo que la FDA (y también cualquier candidato a tomar este medicamento) querría saber. En este sentido, el orden de los grupos también es importante. Te va mejor poner en primer lugar el grupo del fámaco, para que cuando calcules la proporción de sujetos que lo tomaron menos la proporción de sujetos que tomaron el placebo te salga un número positivo si H_a es verdadera. Si cambias los grupos de orden, el signo habría sido negativo.

Ahora calcula el estadístico de contraste:

1. En primer lugar, determinas que $\hat{p}_1 = \frac{26}{374} = 0{,}070$ y $\hat{p}_2 = \frac{8}{210} = 0{,}038$.

   Los tamaños muestrales son $n_1 = 374$ y $n_2 = 210$, respectivamente.

2. La diferencia entre estas proporciones muestrales es $\hat{p}_1 - \hat{p}_2 = 0{,}070 - 0{,}038 = 0{,}032$.

3. La proporción muestral global es $\hat{p} = \frac{26+8}{374+210} = 0{,}058$.

4. El error estándar es $\sqrt{0{,}058(1-0{,}058)\left(\frac{1}{374}+\frac{1}{210}\right)} = 0{,}020$.

5. Por último, el estadístico de contraste es 0,032/0,020=1,60. ¡Buf!

El valor *p* es la probabilidad de alcanzar o superar (en este caso en sentido positivo) 1,60. Esta probabilidad es 1–0,9452=0,0548. Este valor *p* es un poco mayor que 0,05, de manera que, técnicamente, no tienes indicios suficientes para rechazar H_0. Esto significa que, según tus datos, las personas que toman el fármaco no sufren más vómitos que aquellas que toman un placebo.

Una valor *p* muy próximo al mágico pero un tanto arbitrario nivel de significación de 0,05 es lo que en estadística se llama un *resultado marginal*. En el ejemplo anterior, como el valor *p* de 0,0548 está cerca de la frontera entre aceptar y rechazar H_0, generalmente se considera un resultado marginal y debe comunicarse como tal.

Lo bonito de dar a conocer un valor *p* es que eres tú quien toma la decisión final. Cuanto más pequeño es el valor *p*, más indicios tienes en contra de H_0, pero ¿cuándo puedes considerar que tienes suficientes indicios? Cada persona es diferente. Si te encuentras el informe de un estudio donde alguien haya encontrado un resultado estadísticamente significativo y ese resultado sea importante para ti, entérate de cuál era el valor *p* para poder tomar tu propia decisión. (Tienes más información en el capítulo 14.)

Parte V

Estudios estadísticos y la búsqueda de una relación significativa

The 5th Wave — Rich Tennant

—VALE, ANALICEMOS LAS PROBABILIDADES ESTADÍSTICAS DE ESTA SITUACIÓN. NOSOTROS SOMOS CUATRO Y ÉL, UNO. JAIME PROBABLEMENTE SE PONDRÁ A GRITAR. MARÍA PROBABLEMENTE SE DESMAYARÁ. TÚ PROBABLEMENTE TE ENFADARÁS CONMIGO POR DEJARME EL COCHE ABIERTO, Y ES BASTANTE PROBABLE QUE YO ECHE A CORRER COMO UN COBARDE SI VIENE HACIA NOSOTROS.

En esta parte...

Muchas estadísticas que ves y oyes están basadas en los resultados de encuestas, experimentos y estudios de observación. Por desgracia, no puedes creer todo lo que ves ni todo lo que oyes.

En esta parte hablaré sobre lo que ocurre realmente entre bastidores. Te enseñaré cómo se diseñan y realizan estos estudios y cómo se recopilan (o cómo deberían recopilarse) los datos, para que seas capaz de detectar resultados engañosos. También verás lo que necesitas para realizar tu propio estudio de manera correcta y eficaz.

Además, analizaremos los datos de estudios válidos para encontrar relaciones entre dos variables, ya sea mediante tablas de contingencia (si las dos variables son categóricas) o mediante correlación y regresión (si las dos variables son numéricas), y veremos cómo extraer conclusiones correctas y detectar problemas.

Capítulo 16

Encuestas, encuestas y más encuestas

En este capítulo

- Comprender las repercusiones de las encuestas y sondeos
- Saber diseñar y llevar a cabo encuestas y sondeos
- Detectar resultados sesgados e inexactos

En la explosión informativa actual, las encuestas hacen furor. Todo el mundo quiere saber cuál es la opinión del público en general sobre todo tipo de cuestiones, desde el precio de los medicamentos hasta los métodos de impartir disciplina a los niños, pasando por el índice de aprobación del presidente y los índices de audiencia de los programas de televisión. Las encuestas y los sondeos forman parte de nuestra vida; son un vehículo con el que rápidamente obtienes información sobre cómo te sientes, qué piensas y cómo vives tu vida, y son un medio para difundir información sobre temas importantes en muy poco tiempo. Las encuestas ponen de relieve cuestiones controvertidas, sacuden conciencias, difunden puntos de vista, señalan la importancia de un determinado aspecto, y educan o persuaden al público en general.

Los resultados de las encuestas pueden tener una gran repercusión, porque cuando muchas personas oyen decir que un determinado porcentaje de los ciudadanos hacen tal cosa o tal otra, aceptan esos resultados como la verdad, y luego toman decisiones y se forman opiniones basadas en esa información. Sin embargo, la verdad es que muchas encuestas no proporcionan información correcta ni completa, y a veces ni siquiera imparcial.

En este capítulo comento la repercusión de las encuestas y para qué se utilizan, y te doy un paseo entre bastidores para que veas cómo se diseñan y realizan, con el fin de que sepas analizar críticamente sus resultados y aprendas a hacer tú mismo una encuesta de manera correcta. Asimismo, explico cómo interpretar los resultados de las encuestas y cómo detectar información sesgada e inexacta, para que tú mismo puedas determinar qué resultados son fiables y cuáles deben descartarse.

Comprender la importancia de las encuestas

Una encuesta es un instrumento que recopila datos por medio de preguntas y respuestas. Se utiliza para obtener información sobre las opiniones, conductas, perfiles demográficos, estilos de vida y otras características reseñables de la población de interés.

¿Cuál es la diferencia entre un sondeo y una encuesta? Los estadísticos no establecen una distinción clara entre ambas, pero yo diría que un *sondeo* suele ser una encuesta breve que contiene pocas preguntas (a lo mejor así los investigadores consiguen que responda más gente, diciéndoles que es un sondeo en lugar de una encuesta). En cualquier caso, a todos los efectos las encuestas y los sondeos son la misma cosa.

Todos los días entramos en contacto con las encuestas y con sus resultados. En comparación con otros tipos de estudios, por ejemplo, los experimentos médicos, algunas encuestas son relativamente fáciles de realizar. Proporcionan resultados rápidos que a menudo generan titulares interesantes o reportajes atractivos para periódicos y revistas. Casi todos quedamos convencidos por las encuestas porque tenemos la sensación de que los resultados reflejan las opiniones de personas iguales a nosotros (aunque quizá nunca hayamos participado en una). Además, mucha gente disfruta sabiendo lo que otras personas sienten, qué hacen, adónde van y qué les preocupa. Al ver los resultados de una encuesta, la gente siente que de algún modo forma parte de un grupo más grande. Con eso precisamente cuentan los *encuestadores* (las personas que realizan encuestas), y por eso invierten tanto tiempo en hacer encuestas y sondeos y en comunicar los resultados de sus investigaciones.

Llegar hasta la fuente

¿Quién hace encuestas hoy en día? Pues cualquiera que tenga algo que preguntar. Entre los grupos y entidades que realizan o encargan encuestas y comunican los resultados se encuentran los siguientes:

- ✔ Agencias de noticias.
- ✔ Partidos políticos, gobierno y ministerios.
- ✔ El Instituto Nacional de Estadística (INE), el Eurostat y los institutos y servicios estadísticos de las Comunidades autónomas.
- ✔ Empresas profesionales que realizan sondeos de opinión (por ejemplo la organización Gallup, el Centro de Investigaciones Sociológicas [CIS], Metroscopia, Sigma Dos y Nielsen).
- ✔ Periódicos y revistas, programas de televisión y programas de radio.
- ✔ Órganos consultivos (Consejo Superior de Estadística, Comisión Interministerial de Estadística y Comité Interterritorial de Estadística).

✔ Grupos con intereses especiales (por ejemplo, el Banco de España, la Organización de Consumidores y Usuarios, los gremios y colegios profesionales, etc.).

Los peores coches del milenio

En Estados Unidos hay un programa de radio muy popular llamado Car Talk que se emite los sábados por la mañana por l aemisora National Public Radio. Está presentado por Click y Clack dos hermanos de Cambridge, Massachusetts, que en un tono distendido ofrecen consejos a los oyentes que llaman para preguntar por extrañas averías. De vez en cuando, en el sitio web del programa cuelgan encuestas sobre todo tipo de temas divertidos relacionados con los coches, por ejemplo: "¿Qué dice la pegatina que llevas en el parachoques?". Una de las encuestas formuló la siguiente pregunta: "¿Cuál crees que fue el peor coche del milenio?". Miles y miles de personas respondieron con sus votos, aunque, naturalmente, esas personas no representan a todos los estadounidenses que tienen coche. Únicamente representan a los que oyen ese programa de radio, entraron en el sitio web y contestaron a la pregunta de la encuesta. Para que no te quedes en ascuas (sé que te mueres por saberlo), los resultados de la encuesta figuran en la tabla siguiente. Aunque quizá no tengas edad suficiente para recordar algunos de estos vehículos, en Internet encontrarás montones de imágenes y reseñas sobre todos ellos (pero no olvides que estos resultados tan sólo representan las opiniones de los oyentes de Car Talk que se tomaron la molestia de visitar el sitio web y contestar la encuesta). Los porcentajes no suman cien porque los resultados de la tabla representan únicamente los diez modelos que recibieron más votos.

Puesto	*Modelo*	*Porcentaje de votos*
1	Yugo	33,7
2	Chevy Vega	15,8
3	Ford Pinto	12,6
4	AMC Gremlin	8,5
5	Chevy Chevette	7
6	Renault LeCar	4,3
7	Dodge Aspen/Plymouth Volare	4,1
8	Cadillac Cimarron	4
9	Renault Dauphine	3,6
10	Volkswagen (VW) Bus	2,7

✔ Investigadores académicos.

✔ El gobierno.

✔ Cualquier hijo de vecino (hoy en día no cuesta nada hacer una encuesta a través de Internet).

Algunas encuestas son sólo para echarse unas risas, mientras que otras tienen un propósito más serio. Si te piden participar en una encuesta seria o te dan sus resultados, pregunta siempre cuál es la fuente. Los grupos que tengan un interés especial en los resultados deberían contratar a una organización independiente que se encargue de realizar (o al menos revisar) la encuesta, o bien deberían facilitar copias de las preguntas formuladas. Asimismo, estos grupos deberían dar a conocer el modo en que la encuesta fue diseñada y realizada, para que el público pueda tomar una decisión informada sobre la credibilidad de los resultados.

Preguntar sobre temas candentes

Muchas encuestas se refieren a temas de actualidad; al fin y al cabo la oportunidad y la relevancia para el público son dos de las cualidades más atractivas de una encuesta. Aquí tienes varios ejemplos de temas abordados por encuestas recientes, junto con los resultados que se han hecho públicos:

✔ ¿El activismo de los famosos influye en las opiniones políticas de los estadounidenses? (Más del 90% de los estadounidenses dicen que no, según la cadena CBS News.)

✔ ¿Qué porcentaje de estadounidenses han tenido una relación amorosa con un compañero de trabajo? (Según un sitio web de contactos profesionales, la cifra es el 40%.)

✔ ¿Cuántas personas aquejadas de alguna enfermedad o dolencia buscan información médica en Internet? (La respuesta es el 55%, según una revista médica que se publica en Estados Unidos.)

Cuando lees estos resultados de encuestas, ¿piensas directamente en lo que significan para ti en lugar de preguntarte primero si las cifras son

realmente válidas? Algunos de los resultados anteriores son más válidos y precisos que otros, y antes de aceptarlos como dogma de fe deberías preguntarte si son fiables. Los organismos de investigación reconocidos, las empresas que realizan sondeos de opinión a gran escala y las revistas sometidas a *revisiones inter pares* (todos los artículos publicados en la revista han sido revisados por otros especialistas en la materia y cumplen ciertas normas) son fuentes fiables. Si no conoces el grupo que realiza la encuesta y los resultados son importantes para ti, averigua cuál es la fuente.

Repercusión sobre las personas

Si bien algunas encuestas son un mero entretenimiento, hay otras que pueden afectar directamente a tu vida personal o profesional. Estas encuestas relacionadas con cuestiones vitales deben analizarse con detenimiento antes de tomar ninguna medida o decisión importante. Las encuestas pueden hacer que los políticos modifiquen o promulguen nuevas leyes, motivar a los investigadores para que trabajen en problemas actuales, animar a los fabricantes a inventar nuevos productos o cambiar sus políticas y prácticas empresariales, e influir en el comportamiento y la forma de pensar de la gente. Aquí tienes, a modo de ejemplo, algunos resultados de encuestas que pueden influir en tu vida:

- ✔ **Deterioro de la atención sanitaria infantil.** Según una encuesta a 400 pediatras realizada por el Centro Médico Infantil de Washington, en Estados Unidos, los pediatras dedican una media de entre ocho y doce minutos a cada paciente.
- ✔ **Los adolescentes beben más.** Según un estudio de seguimiento realizado en 2009 por la organización Partnership for a DrugFree America (una entidad estadounidense sin ánimo de lucro que trabaja en programas de prevención de consumo de droga y alcohol entre adolescentes y jóvenes), el número de adolescentes de entre catorce y diecisiete años que consumen alcohol ha aumentado el 4% (del 35% en 2008 al 39% en 2009), lo cual supone una inversión de la tendencia a la baja que venía registrándose durante los diez años anteriores a la encuesta.
- ✔ Fíjate siempre en cómo los investigadores definen los términos que utilizan para recopilar los datos. En el ejemplo anterior, ¿cómo definen "consumo de alcohol"? ¿Vale con que el adolescente haya probado el alcohol una vez? ¿Significa que beben alcohol con cierta frecuencia? Los resultados pueden ser engañosos si el alcance de lo que se está contando es demasiado grande. Preocúpate de averiguar qué preguntas se formularon cuando se recopilaron los datos.
- ✔ **Delitos no denunciados.** La encuesta nacional sobre víctimas de la delincuencia, realizada por la Oficina de Justicia de Estados Unidos, concluye que tan sólo se denuncian a la policía el 49,4% de los delitos violentos. Las razones que aportan las víctimas para no denunciar los delitos a la policía se indican en la tabla 161.

La razón que más veces se adujo para no haber denunciado un delito violento a la policía fue que la víctima lo consideró un asunto personal (19,2%). Observa que casi el 12% de las razones están relacionadas con la percepción del propio proceso de denuncia (por ejemplo, que llevaría demasiado tiempo o que la policía podría molestarse o no resolvería nada).

Tabla 16-1. Razones de que las víctimas no denunciaran delitos violentos

Razones de no denunciar	*Porcentaje de víctimas*
Lo consideró un asunto personal	19,2
El delincuente no llegó a cometer el delito	15,9
Lo denunció a otro funcionario público	14,7
Consideró que el delito no era suficientemente importante	5,5
Pensó que sería molestar innecesariamente a la policía	5,3
Falta de pruebas	5,0
Temor a represalias	4,6
Demasiados inconvenientes / Falta de tiempo	3,9
Pensó que la denuncia no serviría de nada	2,7
La propiedad robada no tenía número de identificación	0,5
No descubrió el delito hasta más tarde	0,4
Otras razones	22,3

Por cierto, ¿te has fijado en lo grande que es la categoría "Otras razones"? Este porcentaje tan grande sin explicar indica que la encuesta podría haber sido más concreta o que pueden hacerse más indagaciones en torno a la razón de que las víctimas no denuncien los delitos. Puede que las propias víctimas no estén seguras.

Entre bastidores: los entresijos de las encuestas

Las encuestas y sus resultados forman parte de tu experiencia cotidiana, y esos resultados te ayudan a tomar decisiones importantes (algunas decisiones pueden incluso dar un giro radical a tu vida). Por eso es fundamental ver las encuestas de forma crítica. Antes de tomar alguna medida o decisión basada en los resultados de una encuesta, debes determinar si esos resultados son verosímiles y fiables. Una buena manera de comenzar a desarrollar estas habilidades detectivescas es meterse entre bastidores y conocer la forma de diseñar, desarrollar, llevar a cabo y analizar una encuesta.

El proceso de una encuesta consta de diez pasos:

1. **Definir el propósito de la encuesta.**
2. **Definir la población objetivo.**
3. **Elegir el tipo de encuesta y el momento de su realización.**
4. **Diseñar la introducción atendiendo a criterios éticos.**
5. **Formular las preguntas.**
6. **Seleccionar la muestra.**
7. **Realizar la encuesta.**
8. **Insistir, insistir e insistir.**
9. **Organizar y analizar los datos.**
10. **Extraer conclusiones.**

Cada paso entraña una serie de dificultades, pero todos ellos son de vital importancia para obtener unos resultados imparciales y exactos. Esta secuencia de pasos te ayudará a diseñar, planificar y llevar a cabo una encuesta, pero también puedes utilizarla para valorar de forma crítica la encuesta de otra persona, si esos resultados son importantes para ti.

Planificar y diseñar una encuesta

El propósito de una encuesta es responder a preguntas sobre una población objetivo. La *población objetivo* es el conjunto de personas sobre las cuales quieres extraer conclusiones. En la mayoría de las situaciones, preguntar a la población entera (es decir, realizar un *censo* en toda regla) es imposible porque para ello los investigadores tendrían que invertir demasiado tiempo o demasiado dinero. Generalmente lo mejor que puedes hacer es seleccionar una muestra de personas de la población objetivo, preguntar a esas personas y luego extraer conclusiones sobre la población entera a partir de los datos de esa muestra.

Parece fácil, ¿verdad? Pues no lo es. Una vez que te das cuenta de que no puedes preguntar a toda la población objetivo, empiezan a surgir los problemas. Después de seleccionar la muestra, muchos investigadores no están seguros de lo que deben hacer para obtener los datos que necesitan. Por desgracia, muchas encuestas se llevan a cabo sin haber pensado detenidamente en estas cuestiones, lo cual se traduce en errores, resultados engañosos y conclusiones erróneas. En los siguientes apartados encontrarás información más detallada sobre los cinco primeros pasos del proceso de encuesta.

Definir el propósito de la encuesta

Quizá esto te parezca una obviedad, pero la realidad es que se han diseñado y realizado muchas encuestas que luego no han cumplido su propósito, o que han cumplido sólo algunos de sus objetivos, pero no todos. Es fácil perderse entre las preguntas y olvidar lo que realmente quieres averiguar. Cuando formules el propósito de una encuesta, procura ser lo más concreto posible. Piensa en las conclusiones que querrías extraer si fueras a elaborar un informe, y guíate por eso para fijar tus objetivos.

A muchos investigadores los árboles no les dejan ver el bosque. Si el dueño de un restaurante quiere determinar el índice de satisfacción de sus clientes, previamente tendrá que pensar en qué tipo de comparaciones quiere realizar y qué información exacta quiere obtener. Tendrá que plantear preguntas referidas al momento en que los clientes entraron en el restaurante (fecha y hora), o incluso a la mesa donde se sentaron. Y si quiere comparar el índice de satisfacción de los adultos con el de los grupos familiares, por poner un ejemplo, tendrá que preguntar cuántas personas había en el grupo y cuántos eran niños. Pero si se limita a hacer un par de preguntas sobre el grado de satisfacción, o bien plantea todas las preguntas que se le ocurran sin pensar antes para qué necesita la información, puede que termine con más preguntas que respuestas.

Cuanto más concreto seas en cuanto al propósito de la encuesta, más fácil te resultará diseñar preguntas que cumplan tus objetivos, y mejor preparado estarás para redactar tu informe.

Definir la población objetivo

Imagina que quieres realizar una encuesta para determinar qué proporción de empleados utilizan el correo electrónico en el lugar de trabajo

para enviar y recibir mensajes personales. Quizá creas que la población objetivo son los empleados que utilizan el correo electrónico con fines profesionales. Sin embargo, lo que quieres saber es la proporción de empleados que lo utilizan para mensajes personales, de manera que no puedes preguntarles sólo a los que utilizan el correo electrónico para trabajar, porque entonces los resultados estarán sesgados contra las personas que no envían ni reciben mensajes de correo electrónico en el lugar de trabajo. Pero entonces, ¿deberías contar también con los trabajadores que no tienen siquiera acceso a un ordenador durante su jornada laboral? (ya ves lo rápido que se complican las cosas).

Yo diría que la población objetivo más adecuada en este caso son todas las personas que utilizan ordenadores conectados a Internet en el lugar de trabajo. Todos los integrantes de este grupo tienen al menos acceso al correo electrónico, aunque sólo algunos de ellos realmente lo utilicen con fines profesionales, y de estos, sólo algunos lo utilicen para enviar y recibir mensajes personales (y eso es justo lo que quieres averiguar, cuántos utilizan el correo electrónico para ese fin).

Tienes que definir claramente la población objetivo. Tu definición es lo que te ayudará a seleccionar una muestra adecuada y te orientará en tus conclusiones, para que no te excedas al extrapolar los resultados. Si el investigador no ha definido claramente la población objetivo, es probable que surjan otros problemas con la encuesta.

Elegir el tipo de encuesta y el momento de su realización

El siguiente paso consiste en decidir cuál es el tipo de encuesta más apropiado para la situación que tienes entre manos. Las encuestas se pueden hacer por teléfono, por correo, con entrevistas de puerta en puerta o a través de Internet. Sin embargo, no todos los tipos de encuestas son adecuados para todas las situaciones. Por ejemplo, pongamos que quieres determinar algunos de los factores que están relacionados con el analfabetismo en tu país. No puedes enviar una encuesta por correo porque entonces las personas que no sepan leer no podrán contestarla. En ese caso será más apropiado hacer una encuesta telefónica.

Elige el tipo de encuesta que sea más adecuado para la población objetivo, es decir, el que te permita obtener los datos más fidedignos y significativos que sea posible. Ten presente también tu presupuesto: las entrevistas de puerta en puerta son más caras que las encuestas telefónicas, por ejemplo. Cuando analices los resultados de una encuesta, comprueba si el tipo de encuesta utilizado es el más apropiado para la situación, teniendo en cuenta los límites del presupuesto.

A continuación, tienes que decidir cuándo realizar la encuesta. En la vida es fundamental elegir siempre el momento adecuado, y lo mismo ocurre con las encuestas. Los sucesos actuales influyen en las opiniones de la gente, y aunque algunos encuestadores intenten averiguar cuál es el sentir real de los encuestados, hay otros que se aprovechan de esos sucesos, especialmente de los negativos, y los utilizan como plataformas políticas o como carnaza para generar titulares y controversias. Por ejemplo, después de un tiroteo es habitual que surjan encuestas sobre el control de armas. Asimismo, hay que considerar otros acontecimientos que coincidan en el tiempo con la encuesta; por ejemplo, puede que la gente no quiera contestar el teléfono durante una retransmisión deportiva importante, el día de unas elecciones, durante los juegos olímpicos o en vacaciones. Elegir un mal momento puede ser una fuente de sesgo en los resultados.

Además de la fecha, la hora del día también es importante. Si realizas una encuesta telefónica para conocer la opinión de la gente sobre el estrés en el lugar de trabajo y llamas a una casa entre las nueve de la mañana y las cinco de la tarde, puedes estar seguro de que tus resultados estarán sesgados, porque ése es el horario en que la mayoría de las personas están trabajando (están ocupados estresándose).

Diseñar la introducción atendiendo a criterios éticos

Si bien esta regla no se aplica a las pequeñas encuestas que se publican en Internet y en las revistas, las encuestas serias deben proporcionar información relacionada con cuestiones éticas importantes. En primer lugar, deben tener lo que los encuestadores llaman un *texto introductorio* donde se indique el propósito de la encuesta, qué se hará con los datos, si la información facilitada por los encuestados será confidencial o anónima (lee el recuadro "Anonimato y confidencialidad" más adelante en este mismo capítulo), y que la participación es de agradecer pero no obligatoria. El texto introductorio también debe contener los datos de contacto del investigador para que los encuestados puedan dirigirse a él en caso de duda.

En la mayoría de los países existe una normativa específica que regula la realización de encuestas y los límites éticos de obtención y uso de los datos que de ellas se obtengan. Asegúrate también de que cumples todos los requisitos que exige la ley en cuanto a protección y almacenaje de datos en tu ámbito de estudio. Además, algunas organizaciones concretas tienen también su normativa interna, que deberás tener en cuenta.

Formular las preguntas

Cuando ya está claro el propósito, el tipo, el momento de realización y los aspectos éticos de la encuesta, el siguiente paso es formular las preguntas. El planteamiento de las preguntas puede suponer una enorme diferencia en cuanto a la calidad de los datos que se recopilen. Una fuente de sesgo muy habitual en las encuestas es la forma en que están redactadas las preguntas. Se ha comprobado que la redacción de las preguntas influye directamente en el resultado de la encuesta. Las *preguntas capciosas*, también llamadas *preguntas tendenciosas*, se diseñan con el propósito de favorecer una determinada respuesta por encima de otra. Estas preguntas pueden influir mucho en los encuestados, hasta el

punto de que sus respuestas quizá no reflejen lo que realmente piensan sobre una cuestión.

Por ejemplo, aquí tienes dos maneras de plantear una pregunta (las dos capciosas) acerca de la propuesta de subir los impuestos para financiar la construcción de nuevas escuelas:

- *¿No está usted de acuerdo en que un pequeño incremento porcentual en el IVA es una inversión que merece la pena para mejorar la calidad de la educación de nuestros hijos?*
- *¿No cree usted que deberíamos frenar el incremento de la carga fiscal de los contribuyentes y abstenernos de pedir otra subida del IVA para recaudar fondos destinados a un sistema de enseñanza que despilfarra sus recursos?*

Por la forma en que están redactadas estas dos preguntas tendenciosas puedes ver fácilmente qué quieren los encuestadores que respondas. Para empeorar las cosas, ninguna de las dos preguntas menciona a cuánto asciende exactamente la subida de impuestos propuesta, lo cual también es engañoso.

La redacción de la pregunta debe ser neutra, de modo que se proporcione al lector la información que necesita para tomar una decisión con conocimiento de causa. Por ejemplo, la pregunta sobre la subida de impuestos estaría mejor formulada de la manera siguiente:

El Ministerio de Educación propone incrementar el 0,01% el IVA con el fin de obtener fondos para construir 43 nuevas escuelas de primaria. ¿Qué opina sobre el impuesto que se propone? (Posibles respuestas: muy a favor, a favor, ni a favor ni en contra, en contra, muy en contra.)

Si el propósito de la encuesta es únicamente recopilar información, y no influir ni persuadir al encuestado, las preguntas deben formularse de manera neutra e informativa con el fin de reducir al mínimo el sesgo. El mejor modo de valorar la neutralidad de una pregunta consiste en preguntarte a ti mismo si sabes lo que el encuestador quiere que contestes. Si la respuesta es sí, la pregunta es capciosa y puede conducir a resultados engañosos.

Si los resultados de una encuesta son importantes para ti, pídele al investigador una copia de las preguntas utilizadas a fin de valorar su calidad. Si eres tú quien realiza la encuesta, haz que otras personas revisen las preguntas para estar seguro de que las has redactado de forma neutra e informativa.

Seleccionar la muestra

Una vez diseñada la encuesta, el siguiente paso es seleccionar las personas que van a participar en ella. Como generalmente no tienes tiempo ni dinero suficiente para realizar un censo (una encuesta de toda la población objetivo), tienes que seleccionar un subconjunto de la población, lo que se llama una *muestra*. La manera de seleccionar esta muestra puede influir enormemente en la precisión y la calidad de los resultados.

Existen tres criterios importantes para seleccionar una muestra válida, como verás en los siguientes apartados:

Una muestra válida representa a la población objetivo

Para que represente a la población objetivo, la muestra debe seleccionarse de dicha población objetivo, sin excluir ni añadir a otras personas. Supongamos que quieres averiguar cuántas horas diarias, en promedio, ven la televisión los habitantes de una ciudad. En este caso no bastaría con preguntar a los estudiantes de una residencia universitaria por sus hábitos televisivos, porque representan tan sólo a una parte de la población objetivo.

Por desgracia, muchas personas que hacen encuestas no dedican el tiempo ni el dinero necesario a seleccionar una muestra representativa de personas que participen en el estudio, de modo que terminan obteniendo resultados sesgados. Cuando caigan en tus manos los resultados de una encuesta, averigua cómo fue seleccionada la muestra y hasta qué punto dichos resultados coinciden con los de la población objetivo.

Una muestra válida se selecciona aleatoriamente

Una muestra *aleatoria* es aquella que tiene la misma probabilidad de ser seleccionada que cualquier otra muestra posible (de igual tamaño) de la población objetivo. La manera más fácil de entenderlo es imaginar un sombrero (o un cubo) lleno de tiras de papel, cada una con el nombre de una persona escrito sobre ella; si las tiras se mezclan bien antes de extraerlas de una en una, el resultado será una muestra aleatoria de la población objetivo (en este caso, la población de personas cuyos nombres están en el sombrero). Una muestra aleatoria elimina el sesgo en el proceso de muestreo.

Las empresas que realizan sondeos de opinión y son reconocidas por hacerlo bien, como por ejemplo Gallup, utilizan un procedimiento de marcación aleatoria de dígitos para llamar por teléfono a los integrantes de su muestra. Naturalmente, esto excluye a las personas que no tienen

teléfono, pero como hoy en día la inmensa mayoría de las viviendas tienen teléfono, el sesgo generado por la exclusión de las personas que no tienen teléfono es relativamente pequeño.

Mucho ojo con las encuestas que tienen una muestra muy grande pero que no se ha seleccionado de forma aleatoria. Las encuestas a través de Internet son un claro ejemplo. Si 50.000 personas visitan un sitio web y responden a una encuesta, el administrador de ese sitio conseguirá un montón de datos. Sin embargo, la información estará sesgada: se ha comprobado que las personas que responden a las encuestas suelen tener opiniones más tajantes que quienes no responden. Y si de entrada los participantes no se han seleccionado aleatoriamente, ya puedes imaginar hasta qué punto las opiniones de los encuestados serán tajantes (y estarán sesgadas). Si el diseñador de la encuesta tomara una muestra más pequeña pero aleatoria, los resultados de la encuesta serían más precisos.

Una muestra válida debe ser suficientemente grande para que los resultados sean precisos

Si tienes un tamaño muestral grande y además la muestra es representativa de la población objetivo y se ha seleccionado de forma aleatoria, puedes estar seguro de que la información obtenida será bastante precisa. El grado de precisión depende del tamaño muestral: cuanto más grande sea la muestra, más precisa será la información (siempre y cuando dicha información sea válida). La precisión de las preguntas de una encuesta se mide mediante un porcentaje. Este porcentaje, llamado *margen de error*, representa cuánto supone el investigador que los resultados podrían variar si repitiera la encuesta muchas veces utilizando diferentes muestras del mismo tamaño. Encontrarás más información al respecto en el capítulo 12.

Una manera rápida de conocer de forma aproximada la precisión mínima de una encuesta que maneje datos categóricos (por ejemplo el sexo o la filiación política) consiste en dividir 1 por la raíz cuadrada del tamaño muestral. Por ejemplo, una encuesta de 1.000 personas (seleccionadas aleatoriamente) tiene una precisión de ± 0,032, es decir, 3,2 puntos porcentuales. (En el capítulo 12 encontrarás la fórmula exacta para calcular la precisión de una encuesta.) En los casos en que no haya respondido todo el mundo, el tamaño muestral debe sustituirse por el número de personas que sí hayan respondido. (Más información en el apartado "Insistir, insistir e insistir".) Recuerda que estas estimaciones de precisión son conservadoras. Si aplicas la fórmula, obtendrás un valor mucho más exacto. (Más información en el capítulo 13.)

Cuando la población es muy grande (del orden de varios miles, por ejemplo), lo que importa es el tamaño de la muestra, no el tamaño de la población. Por ejemplo, si la población es grande y seleccionas una muestra aleatoria de 1.000 personas, tu grado de precisión será de aproximadamente 3,2 puntos porcentuales, con independencia de si la muestra procede de un pueblo de 10.000 habitantes, de una provincia de 1.000.000 habitantes o de todo el país. Esto fue una de las cosas que más me sorprendió cuando empecé a estudiar estadística, y todavía hoy me produce estupefaccción: parece mentira que se pueda conseguir semejante precisión con un tamaño muestral comparativamente tan pequeño.

Sin embargo, cuando la población es pequeña es necesario aplicar otros métodos para determinar la precisión y el tamaño muestral. Por ejemplo, una muestra de 10 personas pertenecientes a una población de 100 representa una parte mayor del total que una muestra de 10 personas pertenecientes a una población de 10.000. Existen métodos más avanzados que incluyen una corrección para poblaciones finitas y que resuelven los problemas asociados a poblaciones pequeñas.

Realizar la encuesta

La fase de diseño ha concluido y los participantes han sido seleccionados. Ahora es el momento de llevar a cabo la encuesta, otro paso importante que puede dar lugar a muchos errores y sesgo.

Recopilar los datos

Durante la encuesta propiamente dicha puede ocurrir que los participantes no entiendan bien las preguntas, que den respuestas no incluidas entre las opciones disponibles (en el caso de preguntas de respuesta múltiple) o que decidan dar respuestas imprecisas o descaradamente falsas; esto último se llama *sesgo de respuesta* (como ejemplo de sesgo de respuesta, piensa en lo difícil que sería obtener respuestas sinceras al preguntar a la gente si alguna vez han hecho trampa en la declaración de la renta).

Algunos de los problemas potenciales asociados al proceso de recopilación de datos pueden reducirse al mínimo o evitarse por completo formando adecuadamente al personal encargado de llevar a cabo la encuesta. Con una formación adecuada es posible resolver de manera coherente y clara cualquier problema que surja durante la encuesta, y además se cometen menos errores al registrar los datos. Los problemas derivados de preguntas confusas o falta de opciones de respuesta pueden resolverse realizando un estudio preliminar con unos pocos participantes antes de la encuesta definitiva, para luego, a partir de los comentarios y sugerencias de esas personas, solucionar cualquier

problema que pueda haber en las preguntas.

Anonimato y confidencialidad

Si tuvieras que hacer una encuesta para determinar qué proporción de empleados utilizan el correo electrónico para mensajes personales en el lugar de trabajo, la tasa de respuesta probablemente sería un problema, ya que muchas personas se mostrarían reacias a hablar sobre esta cuestión, al menos de forma sincera. Una manera de animarlos a responder sería decirles que su privacidad estaría protegida durante y después de la encuesta.

Cuando presentas los resultados de una encuesta, generalmente no vinculas la información recopilada a los nombres de los encuestados, ya que eso sería una vulneración de su privacidad. Probablemente hayas oído hablar antes de los términos anónimo y confidencial, pero quizá no tengas claro que estas dos palabras tienen un significado completamente diferente. Garantizar la confidencialidad significa que podría relacionar tu información con tu nombre en mi informe, pero me comprometo a no hacerlo. Garantizar el anonimato significa que no tengo manera de relacionar tu información con tu nombre en mi informe, aunque quisiera.

Si te piden que participes en una encuesta, entérate bien de lo que tienen previsto hacer los investigadores con tus respuestas y si tu nombre podrá o no relacionarse con la encuesta (las buenas encuestas siempre te dejan este punto completamente claro). Luego decide si todavía quieres participar.

El personal también puede aprender a crear un entorno en el que el encuestado se sienta suficientemente seguro como para decir la verdad. En este sentido, garantizar la protección de la privacidad hace que más gente se anime a responder. A fin de reducir al mínimo el sesgo, el entrevistador debe seguir un guión idéntico para todos los sujetos.

Cuidado con los conflictos de interés generados por las encuestas engañosas. Por ejemplo, si la persona que te ha atendido te pregunta por la calidad del servicio, es posible que no respondas con sinceridad. O si el último día de tratamiento, justo antes de irte, tu fisioterapeuta te pide que rellenes una encuesta "anónima" para conocer tu grado de satisfacción, puede que tenga problemas de sesgo.

Insistir, insistir e insistir

Cualquiera que alguna vez haya tirado una encuesta a la basura o se haya negado a responder "unas pocas preguntas" por teléfono sabe que conseguir que la gente participe en una encuesta no es tarea sencilla. Si el investigador quiere reducir el sesgo al mínimo, lo mejor es conseguir que participen tantas personas como sea posible, y para eso hay que insistir y ponerlo fácil. Ofrece dinero, cupones, sobres franqueados con la dirección ya escrita, posibilidad de ganar premios, etc. Todo ayuda.

Si únicamente contestan las personas sensibilizadas con el tema de la encuesta, el resultado es que tan sólo se tendrán en cuenta sus opiniones, ya que el resto de las personas a quienes no les preocupe demasiado el tema no habrán respondido, y los votos "me da lo mismo" no se habrán contabilizado. O puede que sí les importara pero no se tomaron la molestia de decírselo a nadie. En cualquier caso, su voto no se cuenta.

Por ejemplo, supongamos que 1.000 personas reciben una encuesta en la que se les pregunta si habría que cambiar el reglamento para que se pudiera sacar a pasear a los perros por el parque sin ponerles correa. Lo más probable es que sólo contestaran las personas que están muy a favor o muy en contra de la propuesta. Supongamos que contestan 200 personas, 100 en contra y 100 a favor. Eso significa que se han dejado de contabilizar 800 opiniones. Supongamos que a esas 800 personas les trae sin cuidado esta cuestión. Si pudieras contabilizar sus opiniones, los resultados serían 800/1.000=80% de no sabe/no contesta, 100/1.000=10% a favor del nuevo reglamento, y 100/1.000=10% en contra del nuevo reglamento. Pero sin los votos de esas 800 personas, los investigadores dirían lo siguiente: que de las personas que respondieron, el 50% votó a favor del nuevo reglamento y el 50% votó en contra. Como puedes ver, este resultado (totalmente sesgado) es muy diferente del que habrías obtenido si te hubieran respondido las 1.000 personas.

La *tasa de respuesta* de una encuesta es un porcentaje que se determina dividiendo el número de personas que han respondido por el número de personas cuya participación se ha solicitado. Naturalmente, siempre se busca la tasa de respuesta más alta que sea posible, pero ¿hasta dónde hay que llegar para reducir al mínimo el sesgo? Los estadísticos más puristas consideran que una buena tasa de respuesta debe estar por encima del 70%, pero yo creo que hay que ser un poco más realistas. Actualmente se hacen encuestas a troche y moche, y muchas tasas de respuesta, si no la mayoría, están muy por debajo del 70%. De hecho, es más probable que oscilen entre el 20 y el 30%, a menos que la encuesta la realice una organización profesional o te regalen un coche nuevo por participar.

Cuando examines los resultados de una encuesta, fíjate siempre en la tasa de respuesta. Si es muy baja (muy inferior al 50%), es probable que los resultados estén sesgados, en cuyo caso deben cogerse con pinzas o incluso descartarse.

No te dejes impresionar por el hecho de que una encuesta haya contado con un elevado número de participantes; si la tasa de respuesta es baja, por muchos que hayan respondido, habrá muchos más que fueron preguntados y no respondieron.

Las fórmulas estadísticas de este nivel (entre ellas las fórmulas que hay en este libro) presuponen que el tamaño muestral es igual al número de personas que responden. De esta forma, los estadísticos quieren hacerte ver hasta qué punto es importante insistir para no terminar con datos sesgados debido a una baja tasa de respuesta. No obstante, lo cierto es que los estadísticos saben perfectamente que, por mucho que te esfuerces, no siempre puedes conseguir que todo el mundo responda; de hecho, ni siquiera el censo nacional oficial tiene una tasa de respuesta del 100%. Una manera de combatir este problema después de haber recopilado los datos consiste en desglosarlos para ver hasta qué punto concuerdan con la población objetivo. Si la concordancia es bastante buena, el problema del sesgo es menos grave.

Entonces, ¿qué número hay que poner como *n* en todas esas fórmulas estadísticas tan utilizadas (como la media muestral del capítulo 5)? En lugar del tamaño muestral inicial (el número de personas con las que te has puesto en contacto), tienes que utilizar el tamaño muestral final (el número de personas que han respondido). En los medios de comunicación lo normal es que únicamente mencionen el número de participantes en la encuesta, pero ya has visto que para poder valorar los resultados de forma crítica necesitas conocer también la tasa de respuesta (o el número total de personas preguntadas).

En lo que se refiere a la calidad de los resultados, seleccionar un tamaño muestral inicial pequeño e insistir mucho para que participe el mayor número posible de personas es un planteamiento mucho mejor que seleccionar un grupo muy grande y luego obtener una tasa de respuesta baja, ya que en este segundo caso el sesgo será muy alto.

Interpretar los resultados y descubrir problemas

El propósito de una encuesta es obtener información sobre la población objetivo. Esta información puede incluir opiniones, perfiles demográficos, estilos de vida o comportamientos. Si la encuesta se ha diseñado y realizado de manera imparcial y exacta, y sin perder de vista sus objetivos, los datos recopilados deberían proporcionar información válida sobre la población objetivo (dentro del margen de error indicado; que explico en el capítulo 12). A continuación hay que organizar los datos para obtener una imagen clara de lo que está ocurriendo; realizar un análisis en busca de vínculos, diferencias u otras relaciones de interés; y por último extraer conclusiones basadas en los resultados.

Organizar y analizar

Tras haber realizado la encuesta, llega el momento de organizar los datos y analizarlos (en otras palabras: hacer unos cuantos cálculos y dibujar algunos gráficos). Existen muchas formas de representación gráfica y estadísticas resumidas que pueden crearse y calcularse a partir de los datos de una encuesta en función del tipo de información que se haya recopilado (los datos numéricos, por ejemplo los ingresos, poseen unas características particulares y generalmente se presentan de manera diferente a los datos categóricos, por ejemplo el sexo de la persona). En los capítulos 5 a 7 encontrarás más información sobre el modo de organizar y resumir los datos. En función de la pregunta que haya dado pie a la investigación, es posible realizar distintos tipos de análisis con los datos, entre ellos estimaciones de parámetros poblacionales, contrastes de hipótesis sobre la población o búsqueda de relaciones, por nombrar algunos. En los capítulos 13, 14, 15, 18 y 19 encontrarás más información sobre todos estos tipos de análisis.

Cuidado con las estadísticas y los gráficos engañosos. Los datos de las encuestas no siempre se organizan y analizan de forma imparcial y correcta. En el capítulo 3 puedes consultar por qué pueden estar equivocadas las estadísticas.

Extraer conclusiones

Las conclusiones son la mejor parte de cualquier encuesta, la razón última por la cual los investigadores hacen todo ese trabajo. Si la encuesta se ha diseñado y realizado de la forma debida (se ha seleccionado una muestra válida y los datos se han organizado y resumido correctamente), los resultados deberían representar de manera imparcial y exacta la realidad de la población objetivo. Sin embargo, no todas las encuestas se hacen bien. E incluso cuando una encuesta se hace correctamente, los investigadores pueden malinterpretar o exagerar los resultados, en cuyo caso acaban diciendo más de lo que deberían.

¿Conoces el dicho "ver para creer"? Algunos investigadores son culpables de lo contrario, que sería "creer para ver". Dicho de otro modo, aseguran haber visto lo que querían creer sobre los resultados. Razón de más para que sepas dónde trazar la línea entre conclusiones razonables y resultados engañosos, y puedas ver cuándo otra persona ha cruzado esa línea.

Éstos son algunos errores comunes que se cometen al extraer conclusiones de encuestas:

- ✔ Extrapolar los resultados a una población más grande que la representada realmente por el estudio.

✔ Afirmar que existe una diferencia entre dos grupos cuando en realidad no existe. (Consulta el capítulo 15.)

✔ Decir que los resultados no son científicos y luego presentarlos como si lo fueran.

Para evitar estos errores comunes al extraer conclusiones, haz lo siguiente:

1. **Asegúrate de que la muestra se ha seleccionado correctamente y de que las conclusiones no van más allá de la población representada por esa muestra.**

2. **Lee las salvedades y aclaraciones sobre la encuesta, si las hay, antes de leer los resultados.**

   De este modo, si los resultados no están basados en una encuesta científica (una encuesta exacta y no sesgada), será menos probable que te veas influido por ellos. Podrás juzgar por ti mismo si los resultados de la encuesta son fiables.

3. **Ten cuidado con las conclusiones estadísticamente incorrectas.**

   Si, como parte de los resultados de una encuesta, alguien afirma que existe una diferencia entre dos grupos, asegúrate de que esa diferencia es más grande que el margen de error. Si la diferencia está dentro del margen de error, cabe esperar que los resultados muestrales varíen en esa medida por puro azar, y entonces esa supuesta "diferencia" no se puede extrapolar a toda la población. (Encontrarás más información al respecto en el capítulo 14.)

Debes conocer las limitaciones inherentes a las encuestas y desconfiar de cualquier información procedente de una encuesta en la que no se hayan respetado dichas limitaciones. Las malas encuestas son baratas y fáciles de hacer, pero lo que te dan está acorde con su coste. De todos modos, tampoco te dejes impresionar por las encuestas grandes y caras, porque también pueden tener sesgo. Antes de examinar los resultados de una encuesta, aplica los criterios y consejos de este capítulo para averiguar cómo fue diseñada y realizada. Así podrás juzgar la calidad de los resultados y sabrás a ciencia cierta si se ha cometido algún error.

Capítulo 17

Experimentos: ¿logros médicos extraordinarios o resultados engañosos?

En este capítulo

- Distinguir entre experimentos y estudios de observación
- Analizar los criterios para un experimento válido
- Detectar resultados engañosos

Los avances en medicina llegan y desaparecen en un santiamén. Un día lees en el periódico que han descubierto un tratamiento nuevo y prometedor para curar una enfermedad, y luego te enteras de que el fármaco en cuestión no cumplió las expectativas en la última fase de pruebas. Las compañías farmacéuticas bombardean a los televidentes con anuncios de sus productos y consiguen que millones de personas acudan a sus médicos de cabecera suplicando curas milagrosas para sus enfermedades, a veces sin saber siquiera para qué son los medicamentos. Todo el mundo puede buscar en Internet los detalles de cualquier dolencia, enfermedad o síntoma imaginable, y obtener toda clase de información y consejos. Pero ¿cuánta de esa información es realmente creíble? ¿Y cómo puedes decidir qué es lo que más te conviene si te pones enfermo, tienen que operarte o te encuentras en una emergencia?

En este capítulo haremos un recorrido entre bastidores para conocer los entresijos de los experimentos, la fuerza impulsora de los estudios médicos y de otras investigaciones en las cuales se realizan comparaciones (comparaciones que intentan dilucidar, por ejemplo, qué materiales de construcción son mejores, qué refrescos gustan más a los adolescentes, etc.). Conocerás la diferencia entre experimentos y estudios de observación y descubrirás qué utilidad tienen para ti los experimentos, cuál es la forma correcta de llevarlos a cabo y cómo detectar resultados engañosos. Con tantos titulares sensacionalistas y tantos "expertos" ofreciéndote consejos en todas partes, tienes que utilizar toda tu capacidad crítica para evaluar esa información y los conflictos que a menudo se presentan.

Consideraciones básicas sobre los estudios

Básicamente existen dos tipos de estudios: los experimentos y los estudios de observación. En este apartado veremos qué es exactamente lo que diferencia a los experimentos de otros estudios. Pero antes de entrar en detalles tengo que enseñarte la jerga básica de este campo.

Terminología técnica

Para entender los estudios, primero debes conocer el significado de los términos más utilizados:

- ✔ **Sujetos.** Las personas que participan en el estudio.
- ✔ **Estudio de observación.** Un estudio en el cual el investigador se limita a observar a los sujetos y tomar nota de todo. No interviene de ninguna forma, no introduce cambios y no impone limitaciones ni controles.
- ✔ **Experimento.** Este estudio no sólo observa los sujetos en su estado natural, sino que además les aplica tratamientos de forma deliberada en el marco de una situación controlada y estudia los efectos de dichos tratamientos en los resultados.
- ✔ **Respuesta.** La respuesta es la variable cuyo resultado quiere conocer el investigador. Por ejemplo, si el investigador quiere saber qué le ocurre a la tensión arterial cuando tomas a diario una elevada cantidad de Ibuprofeno, la variable de respuesta es la tensión arterial.
- ✔ **Factor.** Un factor es la variable cuyo efecto sobre la respuesta se está estudiando. Por ejemplo, si quieres saber si un determinado fármaco provoca un aumento de la tensión arterial, el factor es la dosis de fármaco administrada al sujeto. Si quieres saber qué plan de adelgazamiento es el que da mejor resultado, el factor sería el tipo de plan utilizado.

  En un estudio puede haber más de un factor; no obstante, en este libro consideraremos que hay uno solo. Para más información sobre el análisis de estudios de dos factores, incluido el uso del análisis de variancia y las comparaciones múltiples para examinar combinaciones de tratamientos, consulta mi libro *Statistics II for Dummies* (sólo disponible en inglés).
- ✔ **Nivel.** Un nivel es un resultado posible de un factor. Cada factor tiene un determinado número de niveles. En el ejemplo anterior sobre planes de adelgazamiento, el factor es el tipo de plan, y los niveles serían los planes concretos que se estudian (por ejemplo, la dieta de los puntos, la dieta disociativa o la famosa dieta de la patata). Los niveles no tienen por qué ir en aumento; no obstante, en el ejemplo del fármaco los niveles serían las distintas dosis administradas diariamente en cantidades crecientes.
- ✔ **Tratamiento.** Un tratamiento es una combinación de los niveles de los factores que se están estudiando. Si sólo tienes un factor, los niveles y los tratamientos son la misma cosa. Si tienes más de un factor, cada combinación de niveles de los factores recibe el nombre de tratamiento.

  Por ejemplo, si quieres estudiar los efectos del tipo de plan de adelgazamiento y la cantidad de agua ingerida a diario, tienes dos factores: 1) el tipo de plan, con tres niveles (dieta de los puntos, dieta disociativa, dieta de la patata); y 2) la cantidad de agua ingerida, con tres niveles (0,75, 1,5 y 2 litros diarios). En este caso hay $3 \times 3 = 9$ tratamientos: dieta de los puntos y 0,75 litros de agua al día; dieta

de los puntos y 1,5 litros de agua al día, etc. hasta llegar a la famosa dieta de la patata y 2 litros de agua al día. A cada sujeto se le asigna un tratamiento (con la suerte que tengo, seguro que me tocaría el último).

- ✔ **Causa y efecto.** Un factor y una respuesta tienen una relación de causa y efecto si un cambio en el factor tiene como resultado directo un cambio en la respuesta (por ejemplo, un aumento de la ingesta calórica provoca un aumento de peso).

En los siguientes apartados verás las diferencias entre estudios de observación y experimentos, cuándo se utiliza cada uno y cuáles son sus puntos fuertes y débiles.

Observar los estudios de observación

Igual que ocurre con las herramientas, hay que elegir el tipo de estudio adecuado para la tarea que tienes entre manos. Hay ciertas situaciones en que los estudios de observación son el método idóneo. Los estudios de observación más comunes son las *encuestas* y los *sondeos* (se explican en el capítulo 16). Si el objetivo es simplemente averiguar lo que piensa la gente y recopilar algunos datos demográficos (por ejemplo, sexo, edad, ingresos, etc.), las encuestas y sondeos no tienen rival, siempre y cuando se diseñen y se lleven a cabo correctamente.

En otras situaciones, sobre todo cuando se busca una relación de causa y efecto, los estudios de observación no son la mejor elección. Por ejemplo, supongamos que la semana pasada te tomaste un par de comprimidos de vitamina C; ¿fue eso lo que evitó que tus compañeros de trabajo te contagiaran el resfriado? A lo mejor te has salvado porque últimamente duermes un poco más o porque te lavas las manos un poco más a menudo, o puede que esta vez simplemente hayas tenido suerte. Con tantas variables en juego, ¿cómo puedes saber cuál de ellas influyó en el resultado de no haberte resfriado? En este caso necesitas un experimento que tenga en cuenta estas otras variables.

Cuando examines los resultados de un estudio, primero determina cuál era su propósito y si el tipo de estudio es adecuado para tal propósito. Por ejemplo, si se ha realizado un estudio de observación en lugar de un experimento para establecer una relación de causa y efecto, las conclusiones alcanzadas deberían analizarse con gran detenimiento.

Examinar los experimentos

La finalidad de un experimento es ver si la respuesta cambia como resultado del factor que estás estudiando; es decir, buscas una relación de causa y efecto. Por ejemplo, ¿tomar Ibuprofeno provoca un aumento de la tensión arterial? En caso afirmativo, ¿cuál es ese aumento? Pero como los resultados varían en todo experimento, quieres tener la certeza de que tus resultados presentan una alta probabilidad de repetirse, si has descubierto que ocurre algo interesante. Es decir, quieres estar seguro de que tus resultados difícilmente pueden ser fruto del azar (lo que en estadística se conoce como resultados *estadísticamente significativos*). Ése es el objetivo de cualquier estudio, ya sea de observación o experimental.

Un buen experimento se lleva a cabo creando un entorno muy controlado, tanto que el investigador pueda ver si un determinado factor o combinación de factores causa un cambio en la variable de respuesta, y en tal caso, la medida en que ese factor (o combinación de factores) influye en la respuesta. Por ejemplo, a fin de obtener la autorización gubernamental para un fármaco antihipertensivo, los investigadores farmacéuticos realizan experimentos con los que determinan si ese fármaco ayuda a reducir la tensión arterial, qué dosis es la más adecuada para cada población de pacientes, qué efectos secundarios (en su caso) se producen, y en qué medida ocurren esos efectos secundarios en cada población.

Diseñar un experimento válido

La forma de diseñar un experimento es lo que marca la diferencia entre resultados válidos y basura. Como la mayoría de los investigadores procuran redactar comunicados de prensa lo más llamativos posible sobre sus experimentos, tienes que ser capaz de distinguir el grano de la paja para determinar si los resultados que te están contando merecen ser creídos. Para decidir si un experimento es creíble, comprueba si cumple todos los criterios expuestos a continuación. Un experimento válido:

- ✔ **Realiza comparaciones.**
- ✔ **Establece un tamaño muestral suficientemente grande como para que los resultados sean precisos.**
- ✔ **Elige a los sujetos que mejor representan a la población objetivo.**
- ✔ **Asigna los sujetos de forma aleatoria al grupo o grupos experimentales y al grupo de control.**
- ✔ **Comprueba si existen variables de confusión.**
- ✔ **Es ético.**

✔ **Recopila datos válidos.**

✔ **Aplica el análisis de datos apropiado.**

✔ **Extrae conclusiones apropiadas.**

En este apartado te explicaré e ilustraré con ejemplos todos estos criterios.

Diseñar el experimento para realizar comparaciones

Todo experimento debe realizar comparaciones fidedignas para ser creíble. Esto puede parecer obvio, pero a menudo los investigadores están tan deseosos de verificar sus resultados que olvidan (o les da igual) demostrar que la causa real de las diferencias observadas en la respuesta es su factor, y no otros factores, incluida la probabilidad aleatoria.

Por ejemplo, supongamos que un investigador está convencido de que tomar vitamina C previene los resfriados, y para demostrarlo les pide a varias personas que tomen un comprimido de vitamina C al día durante seis meses. Ahora imagina que los sujetos se resfrían muy pocas veces durante ese tiempo. ¿Puede atribuir esos resultados a la vitamina C y a ninguna otra cosa? No; no existe forma de saber si los sujetos habrían estado igual de sanos sin tomar vitamina C, debido a otro(s) factor(es) o por puro azar. No hay nada con lo que comparar los resultados.

Para averiguar cuál es el efecto real (si lo hay) que tu factor tiene sobre la respuesta, necesitas una referencia contra la cual comparar los resultados. Esta referencia es lo que se denomina *grupo de control*. Existen diferentes métodos para crear un grupo de control en un experimento. La situación particular es la que determina cuál de estos métodos debe considerarse el más apropiado. Los tres métodos más habituales para incluir un grupo de control consisten en administrar: 1) un tratamiento ficticio; 2) un tratamiento estándar; o 3) ningún tratamiento. Estos métodos se describen en los siguientes apartados.

Cuando examines los resultados de un experimento, asegúrate de que los investigadores crearon un grupo de control para establecer una referencia. Sin un grupo de control no tienes nada contra lo que comparar los resultados, y entonces no puedes saber si el tratamiento aplicado fue la causa real de las diferencias que puedas haber observado en la respuesta.

Tratamientos ficticios: el efecto placebo

Para el sujeto, un tratamiento ficticio (también llamado *placebo*) es indistinguible de un tratamiento "real". Por ejemplo, cuando se administra un medicamento, los sujetos asignados al placebo reciben una pastilla falsa que tiene exactamente el mismo aspecto y sabor que la pastilla auténtica, pero sólo contiene una sustancia inerte, como el azúcar, en lugar del medicamento real. Un placebo establece una medida de referencia correspondiente a las respuestas que habrían ocurrido de todos modos si no se hubiera administrado el tratamiento estudiado (esto habría sido útil en el estudio de la vitamina C mencionado en el apartado "Diseñar el experimento para realizar comparaciones"). Pero, además, un tratamiento ficticio tiene en cuenta lo que los investigadores llaman *efecto placebo*, una respuesta que la gente experimenta (o cree experimentar) porque saben que están recibiendo algún tipo de "tratamiento" (incluso si ese tratamiento es ficticio, por ejemplo pastillas de azúcar).

Las compañías farmacéuticas deben considerar el efecto placebo al examinar los efectos positivos y negativos de un fármaco. Quizá creas que si el grupo de control toma un placebo, los sujetos no deberían manifestar ningún efecto secundario, pero no es así. Si tú mismo estuvieras tomando una pastilla, supieras que podría tratarse del fármaco auténtico y te preguntaran si estás experimentando o no algún efecto secundario, te sorprendería ver cuál sería tu respuesta.

Si no tienes en cuenta el efecto placebo, tienes que creer que todos los efectos secundarios (o resultados positivos) manifestados por los sujetos se deben al fármaco administrado. En ese caso obtienes un número artificialmente elevado de efectos secundarios, ya que al menos una parte de ellos se deben probablemente al efecto placebo y no al fármaco. Si tienes un grupo de control con el que comparar, coges el porcentaje de personas del grupo experimental que hayan manifestado los efectos secundarios y le restas el porcentaje de personas del grupo de control que hayan manifestado dichos efectos. Lo que te queda es básicamente el número neto de efectos secundarios debidos al fármaco, y no el número bruto de efectos secundarios, algunos de los cuales se deben al efecto placebo.

Se ha comprobado que el efecto placebo es real. Si quieres examinar de manera imparcial los efectos secundarios (o reacciones positivas) de un tratamiento, debes tener en cuenta también los efectos secundarios (o reacciones positivas) que manifieste el grupo de control, sabiendo que esas reacciones se deben exclusivamente al efecto placebo.

Tratamientos estándares

En determinadas situaciones, por ejemplo, cuando los sujetos padecen enfermedades muy graves, la opción de administrar un tratamiento ficticio puede ir en contra de la ética. Un ejemplo famoso de comportamiento no ético ocurrió en 1997. El gobierno de Estados Unidos recibió críticas muy duras por financiar un estudio del VIH con el que se pretendía valorar nuevas dosis de AZT, un fármaco que, según se sabía ya entonces, reducía en dos tercios el riesgo de que las mujeres embarazadas transmitieran el virus al feto. Ese estudio en particular, que contó con la participación de 12.000 mujeres embarazadas de África, Tailandia y la República Dominicana, todas ellas infectadas con el VIH, tenía un diseño mortal. Los investigadores administraron a la mitad de las mujeres distintas dosis de AZT, pero a la otra mitad le dieron pastillas de azúcar. Naturalmente, si el gobierno de Estados Unidos hubiera sabido que se estaba dando un placebo a la mitad de los sujetos, no habría financiado el estudio. No es ético administrar un tratamiento ficticio a alguien que tenga una enfermedad mortal para la cual exista un tratamiento estándar (en este caso, la dosis estándar de AZT).

Cuando el uso de tratamientos ficticios deba descartarse por razones éticas, el nuevo tratamiento se compara con al menos un tratamiento existente o estándar cuya efectividad se haya comprobado. Cuando los investigadores reúnen datos suficientes para ver que uno de los tratamientos funciona mejor que el otro, generalmente detienen el experimento y asignan todos los sujetos al mejor tratamiento, también por razones éticas.

Ningún tratamiento

Si los integrantes del grupo de control no van a recibir "ningún tratamiento", el investigador no tiene otro remedio que comunicárselo, debido a la propia naturaleza del experimento. En este caso no se influye de ningún modo en el comportamiento de los sujetos, pero igualmente sirven como grupo de control al establecer una referencia contra la cual comparar los resultados del grupo o grupos experimentales. Por ejemplo, supongamos que quieres determinar si dar diez vueltas diarias a la manzana caminando rápido disminuye la frecuencia cardíaca en reposo al cabo de seis meses. Los sujetos incluidos en el grupo de control sabrán que ellos no van a dar vueltas a la manzana, ya que, obviamente, eso es algo que no puede simularse (aunque no estaría mal que pudiéramos hacer ejercicio simulado y aun así ponernos en forma, ¿verdad?).

Cuando el grupo de control no recibe ningún tratamiento, igualmente hay que asegurarse de que los grupos de sujetos (en este caso personas que caminan frente a personas que no caminan) son similares en tantos aspectos como sea posible, y que además se cumplen el resto de los criterios para un experimento válido (los tienes en el apartado "Diseñar un experimento válido").

Seleccionar el tamaño muestral

El tamaño de una muestra (válida) influye mucho en la precisión de los resultados. Cuanto más grande es la muestra, más precisos son los resultados y más potentes son las pruebas estadísticas (para poder detectar resultados reales, si los hay). En este apartado comento los aspectos más importantes de esta cuestión, y en el capítulo 14 entro en detalles.

La palabra "muestra" generalmente se asocia a las encuestas, donde se selecciona una muestra aleatoria de la población objetivo (tal como se describe en el capítulo 16). Sin embargo, en el contexto de los experimentos, una muestra es el grupo de sujetos que se han presentado voluntarios para participar.

Muestras pequeñas, conclusiones a pequeña escala

De todos los resultados de investigación que se dan a conocer en los medios y que hacen referencia a poblaciones muy grandes, te sorprendería saber cuántos están basados en muestras muy pequeñas. Esto es motivo de preocupación para los estadísticos, quienes conocen de primera mano la dificultad que entraña detectar resultados estadísticamente significativos en una población grande utilizando una muestra pequeña, ya que los conjuntos de datos pequeños presentan más variabilidad de una muestra a otra (la explicación está en el capítulo 12). Si el tamaño muestral es pequeño y el investigador extrae conclusiones a gran escala, significa que no ha utilizado el contraste de hipótesis correcto para analizar los datos (por ejemplo, utilizar la distribución *Z* en lugar de la distribución *t*; vuelve al capítulo 10) o bien la diferencia era tan grande que sería difícil equivocarse. Esto último, no obstante, no siempre se cumple.

Ojo con los estudios que encuentran resultados significativos a partir de tamaños muestrales pequeños (sobre todo en experimentos que comparan muchos tratamientos pero tienen pocos sujetos asignados a cada tratamiento). A los estadísticos les gusta ver al menos cinco sujetos en cada tratamiento, pero cuantos más, mejor. A pesar de todo, hay que ser consciente de algunas limitaciones propias de los experimentos (por ejemplo, los costes, el tiempo y los problemas éticos) y entender que el número de sujetos de un experimento es a menudo inferior al número de participantes en una encuesta.

Si los resultados son importantes para ti, pide una copia del informe de investigación, averigua qué tipo de análisis han hecho con los datos y comprueba si la muestra de sujetos es realmente representativa de la población sobre la cual los investigadores están extrayendo conclusiones.

Definir el tamaño muestral

Si preguntas por el *tamaño muestral*, tienes que dejar claro a qué te refieres con este término. Por ejemplo, puedes preguntar cuántos sujetos fueron seleccionados para participar o también cuántos terminaron realmente el experimento, ya que estos dos números pueden ser muy diferentes. Asegúrate de que los investigadores tienen una explicación para los casos en que los sujetos decidieran abandonar el experimento o no pudieran terminarlo (por alguna razón).

Por ejemplo, en el periódico *The New York Times* leo un artículo titulado "Marihuana: alivio eficaz para el tratamiento contra el cáncer", en cuyo primer párrafo se afirma que esta droga es "mucho más eficaz" que cualquier fármaco destinado a mitigar los efectos secundarios de la quimioterapia. Cuando entras en detalles, descubres que los resultados se basan en tan sólo 29 pacientes (15 de ellos siguieron el tratamiento, y 14 tomaron un placebo). Y luego te enteras de que únicamente 12 de los 15 pacientes del grupo experimental llegaron a terminar el estudio. ¿Qué pasó con los otros tres sujetos?

A veces los investigadores basan sus conclusiones únicamente en los sujetos que terminaron el estudio. Esto puede inducir a error, ya que la falta de información sobre los sujetos que abandonaron (y por qué) puede conducir a datos sesgados. En el capítulo 13 hablo sobre el tamaño muestral que se necesita para alcanzar cierto grado de precisión.

La precisión no es lo único necesario para que los datos puedan considerarse "válidos". También debes eliminar el sesgo seleccionando una muestra aleatoria (en el capítulo 16 explico la forma de hacerlo).

Seleccionar a los sujetos

Una vez elegido el tamaño muestral, el siguiente paso de un experimento consiste en seleccionar a los sujetos (participantes). Aunque a los investigadores les gustaría que los sujetos se seleccionaran de forma aleatoria de sus respectivas poblaciones, en la mayoría de los casos esto no es adecuado. Por ejemplo, supongamos que un grupo de investigadores oftalmológicos quieren probar un nuevo método de cirugía láser para miopes. Necesitan una muestra aleatoria de sujetos, de manera que seleccionan aleatoriamente a varios oftalmólogos de todo el país y seleccionan aleatoriamente a pacientes miopes de esos oftalmólogos. Llaman por teléfono a cada una de las personas seleccionadas y le dicen: "Estamos experimentando con una nueva técnica de cirugía láser para la miopía, y usted ha sido seleccionado al azar para que participe en nuestro estudio. ¿Cuándo le va bien que lo operemos?". Algo me dice que este enfoque no daría buen resultado con la mayoría de las personas (aunque probablemente algunos estarían encantados de aceptar, sobre todo si tienen una cola de espera de dos años para la intervención).

Obtener una muestra de personas realmente aleatoria para que participen en un experimento generalmente es más difícil que obtener una muestra aleatoria de personas para que participen en una encuesta. No obstante, los estadísticos pueden diseñar el experimento de manera que se reduzca al mínimo la posibilidad de sesgo.

Realizar asignaciones aleatorias

Una manera de reducir al mínimo el sesgo en un experimento es introducir algo de aleatoriedad. Una vez seleccionada la muestra, los sujetos se asignan aleatoriamente al grupo o grupos experimentales y al grupo de control. Los grupos experimentales reciben los distintos tratamientos que se estén estudiando, mientras que el grupo de control recibe el tratamiento actual (o estándar), ningún tratamiento o un placebo (está explicado en el apartado anterior "Diseñar el experimento para realizar comparaciones").

Asignar aleatoriamente los sujetos a los distintos tratamientos es de vital importancia para reducir al mínimo el sesgo en un experimento.

Imagina que un investigador quiere determinar los efectos del ejercicio físico sobre la frecuencia cardíaca. Los sujetos de su grupo experimental salen a correr 8 kilómetros y les miden las pulsaciones antes de empezar y después. Los sujetos de su grupo de control se quedan sentados en el sofá viendo reposiciones de series antiguas. ¿A ti qué grupo te gustaría que te tocara? Siempre hay adictos al deporte que preferirían estar en el grupo experimental, pero, si no te atrae la idea de correr 8 kilómetros, podrías ir a lo fácil y presentarte voluntario para despanzurrarte en el sofá (o también puede que odies tanto las reposiciones que serías capaz de correr 8 kilómetros para evitarlas).

Encontrar voluntarios

Para encontrar sujetos dispuestos a participar en sus experimentos, los investigadores generalmente ponen anuncios y ofrecen incentivos como dinero, tratamientos gratuitos o un seguimiento clínico a cambio de su participación. La investigación médica con personas es compleja, pero también necesaria para saber a ciencia cierta si un tratamiento funciona,cómo de bien funciona, qué dosis es la adecuada y cuáles son los efectos secundarios. A fin de prescribir el tratamiento adecuado en las dosis adecuadas en situaciones reales, los médicos y los pacientes dependen de que estos estudios sean representativos de la población en general. Para contar con la participación de sujetos representativos, los investigadores tienen que hacer una extensa campaña publicitaria y seleccionar a suficientes participantes con suficientes características diferentes como para tener una muestra representativa de la población de personas a quienes se recetará esos tratamientos en el futuro.

¿Qué repercusiones tendría sobre los resultados del estudio el hecho de que los propios participantes eligieran grupo? Si para el grupo experimental se presentaran voluntarios únicamente los adictos al deporte (quienes probablemente tienen una frecuencia cardíaca excelente), el investigador tan sólo analizaría los efectos del tratamiento (correr 8 kilómetros) sobre personas muy saludables y activas. No vería cuáles son los efectos sobre la frecuencia cardíaca de personas muy sedentarias. Esta asignación no aleatoria de los sujetos al grupo experimental y al grupo de control podría influir mucho en las conclusiones que luego extrajera del estudio.

Para evitar el sesgo en los resultados de un experimento, los sujetos deben ser asignados aleatoriamente a los tratamientos por un tercero, y no se les debe permitir que elijan grupo según sus preferencias personales. El objetivo de la asignación aleatoria es crear grupos homogéneos, de manera que en todos los grupos exista la misma probabilidad de sesgo o características inusuales. Ten esto siempre presente cuando evalúes los resultados de un experimento.

Controlar las variables de confusión

Imagina que participas como sujeto en un estudio de investigación que examina los factores que influyen en coger un resfriado. Si el investigador se limita a anotar si te has resfriado después de cierto período de tiempo y te hace preguntas sobre tus hábitos (cuántas veces al día te lavas las manos, cuántas horas duermes cada noche, etc.), estará realizando un estudio de observación. El problema de este tipo de estudio es que, si no se controlan otros factores que puedan ejercer alguna influencia y no se regulan las cosas que haces en cada momento, el investigador no podrá determinar exactamente cuál de tus acciones influyó realmente en el resultado (si es que alguna lo hizo).

La mayor limitación de los estudios de observación es que en realidad no permiten ver auténticas relaciones de causa y efecto, debido a lo que en estadística se llama variables de confusión. Una *variable de confusión* es una variable o factor que no se estaba controlando en el estudio pero que pudo influir en los resultados.

Por ejemplo, hace tiempo se publicó una noticia con el siguiente titular: “Un estudio relaciona la maternidad tardía con una mayor longevidad”. El primer párrafo decía que las mujeres que dan a luz a su primer hijo después de cumplir los cuarenta años tienen una probabilidad mucho más alta de vivir más de cien años, en comparación con las mujeres que tienen su primer hijo siendo más jóvenes. Cuando lees los pormenores del estudio (hecho en 1996), descubres, en primer lugar, que se basó en 78 mujeres que residían en zonas residenciales de las afueras de Boston, nacieron en 1896 y vivieron al menos cien años, en comparación con 54 mujeres que también nacieron en 1896 pero fallecieron en 1969 (el primer año para el cual los investigadores pudieron encontrar registros de mortalidad informatizados). El supuesto “grupo de control” estaba formado por mujeres que vivieron exactamente setenta y tres años, ni uno más ni uno menos. De las mujeres que pasaron de los cien años de edad, el 19% habían dado a luz después de cumplir los cuarenta, mientras que únicamente el 5,5% de las mujeres que fallecieron a los setenta y tres habían dado a luz después de cumplir los cuarenta.

Tengo un problema gordo con estas conclusiones. ¿Cómo puede ser que el “grupo de control” se basara exclusivamente en las mujeres que fallecieron en 1969 a la edad de setenta y tres años? ¿Qué pasa con el resto de las mujeres que fallecieron antes de cumplir los setenta y tres o que fallecieron a una edad comprendida entre setenta y tres y cien años? ¿Qué pasa con el resto de las variables que pudieron influir en la edad de las mujeres al dar a luz o en su esperanza de vida, como por ejemplo la situación económica, la estabilidad matrimonial u otros factores socioeconómicos? Las mujeres de este estudio ya habían cumplido los treinta cuando vivieron la Gran Depresión; eso pudo influir en su longevidad y en el momento elegido para dar a luz.

Los investigadores controlan las variables de confusión lo mejor que pueden, las anticipan en la medida de lo posible y procuran reducir al mínimo su posible efecto sobre la respuesta. En los experimentos con personas, los investigadores se enfrentan a muchas variables de confusión.

Por ejemplo, para un estudio que intente determinar el efecto de diferentes tipos y volúmenes de música sobre el tiempo que pasa una persona comprando en una verdulería (sí, hasta eso se investiga), los investigadores tienen que prever y controlar el mayor número posible de variables de confusión. ¿Qué otros factores, aparte del volumen y el tipo de música, podrían influir en el tiempo que pasa alguien en una verdulería? Se me ocurren varios: el sexo de la persona, la edad, la hora del día, si va con niños o no, cuánto dinero lleva encima, el día de la semana, cómo de limpia está la tienda, lo amables que sean los dependientes, y (lo más importante de todo), por qué razón ha ido (¿quiere hacer la compra de toda la semana o tan sólo ha entrado a por una lechuga?).

¿Cómo pueden los investigadores controlar todos esos factores de confusión? Algunos pueden tenerse en cuenta al diseñar el estudio, por ejemplo, la hora del día, el día de semana y la razón de haber ido a comprar. Sin embargo, hay otros factores (por ejemplo, la percepción de las condiciones de la tienda) que dependen totalmente de la persona que participa en el estudio. La forma de controlar esas variables de confusión específicas de cada cual consiste en utilizar pares de personas emparejadas con arreglo a variables importantes, o bien utilizar a la misma persona dos veces: una vez con el tratamiento y la segunda vez sin él. Se dice entonces que el experimento tiene un *diseño de datos apareados*. (Encontrarás más información al respecto en el capítulo 15.)

Antes de creerte una noticia referida a un avance médico (o cualquier noticia que aporte estadísticas), averigua cómo se llevó a cabo el experimento. Los estudios de observación no pueden controlar las variables de confusión, de manera que sus resultados no tienen tanta significación estadística (con independencia de lo que digan las estadísticas) como los resultados de un experimento bien diseñado. En los casos en que no sea posible realizar un experimento (al fin y al cabo, no se puede obligar a una mujer a tener un hijo antes o después de los cuarenta), asegúrate de que el estudio de observación está basado en una muestra suficientemente grande que sea representativa de la población. Y piensa en posibles variables de confusión que puedan influir en las conclusiones.

Problemas éticos

El problema de los experimentos es que algunos diseños experimentales no son éticos. Por poner un ejemplo, no puedes obligar a los sujetos a que fumen para ver si desarrollan cáncer de pulmón. Tienes que examinar a personas que tengan cáncer de pulmón e ir hacia atrás para determinar qué *factores* (las variables estudiadas) pueden haber causado la enfermedad. Pero como no puedes controlar los diversos factores en los que estás interesado (ni ninguna otra variable, en realidad), en los estudios de observación es muy difícil aislar una causa concreta. Pero eso hicieron falta tantas pruebas para demostrar que fumar causa cáncer de pulmón, y por eso pasaron tantos años hasta que se obligó a las tabaqueras a compensar económicamente a las personas afectadas. Aunque las causas del cáncer y de otras enfermedades no pueden determinarse realizando experimentos con personas (no sería ético), los nuevos tratamientos contra el cáncer sí pueden ser (y están siendo) verificados por medio de experimentos. Los estudios médicos en los que se realizan experimentos se denominan *ensayos clínicos*. El gobierno de Estados Unidos tiene un registro de todos los ensayos clínicos que se realizan en el mundo con fondos públicos o privados, y también dispone de información sobre quiénes pueden participar en varios ensayos clínicos. Encontrarás más información al respecto en www.clinicaltrials.gov.

Los experimentos serios deben superar numerosas pruebas que pueden tardar años en llevarse a cabo. La aprobación de un nuevo medicamento, por ejemplo, es un proceso muy lento y minucioso que, en Estados Unidos se encuentra bajo el control y la supervisión de la FDA (la Food and Drug Administration, el equivalente de la Agencia Española del Medicamento). Una razón de que los medicamentos de venta con receta tengan un precio tan elevado es la enorme cantidad de tiempo y dinero necesarios para la investigación y el desarrollo de nuevos fármacos, la mayoría de los cuales no consiguen superar los ensayos y no consiguen la autorización para ser comercializados.

Además, todos los experimentos donde los sujetos son personas están regulados por la ley y deben contar con la autorización oficial. Las investigaciones que se realizan con animales no están tan bien reguladas y continúan generando mucha controversia en todo el mundo.

Las encuestas, sondeos y otros estudios de observación van bien para conocer la opinión de la gente, para examinar sus estilos de vida sin influir en ellos o para analizar algunas variables demográficas. Si tu propósito es determinar la causa de un determinado resultado o comportamiento (es decir, la razón de que haya ocurrido algo), es mucho mejor realizar un experimento. Si un experimento es inviable por razones éticas (o porque resulta muy caro o por otros motivos), lo siguiente mejor es realizar un número elevado de estudios de observación que analicen muchos factores distintos y lleguen a conclusiones similares. (En el capítulo 18 encontrarás más información sobre las relaciones de causa y efecto.)

Recopilar datos válidos

Para determinar si los datos son válidos, los estadísticos aplican tres criterios, cada uno de los cuales está relacionado con la calidad del instrumento de medición utilizado en el proceso de recopilación de esos datos. Para considerarse válidos, los datos de un estudio deben reunir

las siguientes características:

- ✔ **Los datos son fiables. Puedes obtener resultados reproducibles con mediciones posteriores.** Muchas básculas de baño proporcionan una lectura poco fiable. Te subes y te da un número. No te lo crees, así que te bajas, te vuelves a subir y te da otro número (si el segundo número es más bajo, lo más probable es que ya te conformes; de lo contrario, puede que continúes subiendo y bajando de la báscula hasta que veas un número que te guste). O puedes hacer lo mismo que algunos investigadores: tomar tres mediciones y calcular la media (al menos de este modo aumenta un poco la fiabilidad).

  Los datos poco fiables se deben a instrumentos de medición poco fiables o métodos de recopilación de datos poco fiables. Además de las básculas y similares, los errores pueden darse en instrumentos de medición más intangibles, como por ejemplo las preguntas de encuestas, que pueden proporcionar resultados poco fiables si se redactan de manera ambigua (repasa el capítulo 16).

  Cuando examines los resultados de un estudio, entérate de cómo recopilaron los datos. Si las mediciones son poco fiables, los datos podrían ser inexactos.

- ✔ **Los datos son adecuados. Miden lo que se supone que deben medir.** Para comprobar la adecuación de los datos tienes que dar un paso atrás y ver las cosas con perspectiva. Debes hacerte la siguiente pregunta: ¿estos datos miden lo que tienen que medir? ¿O los investigadores deberían haber recopilado otros datos distintos? La idoneidad del instrumento de medición utilizado también es importante. Por ejemplo, muchos educadores consideran que el expediente de un alumno no es una medida válida de su capacidad para tener un buen rendimiento escolar. La alternativa puede ser un planteamiento más holístico que, además de las notas, tenga en cuenta elementos como el esfuerzo, la creatividad, la participación, las actividades extraescolares, etc.

  Antes de aceptar los resultados de un experimento, averigua qué datos se han medido y de qué modo. Asegúrate de que los investigadores han recopilado datos válidos y apropiados para los objetivos del estudio.

- ✔ **Los datos no están sesgados. No contienen errores sistemáticos que distorsionen los valores reales al alza o a la baja.** Los datos están sesgados cuando presentan una diferencia sistemática respecto del valor verdadero. El sesgo puede ocurrir prácticamente en cualquier momento durante el diseño o la realización de un estudio. El sesgo puede estar causado por un instrumento de medición defectuoso (como esa báscula de baño que siempre marca 5 kilos de más), por preguntas que inducen a los encuestados a decantarse por una determinada respuesta, o por investigadores que saben qué tratamiento ha recibido cada sujeto y tienen expectativas preconcebidas.

El sesgo es probablemente el mayor problema a la hora de recopilar datos válidos. No obstante, es posible reducirlo al mínimo utilizando métodos similares a los que ya vimos en el capítulo 16 para las encuestas y en el apartado "Realizar asignaciones aleatorias" de este mismo capítulo, y también haciendo que el experimento tenga doble enmascaramiento.

Doble enmascaramiento (a menudo llamado "doble ciego") significa que tanto los sujetos como los investigadores desconocen qué tratamiento recibe cada persona y quién está en el grupo de control. Los sujetos no deben saber qué tratamiento se les está administrando porque de otro modo los investigadores no podrían medir el efecto placebo. En cuanto a los investigadores, tampoco deben estar al tanto de esa información para que no traten a los sujetos de manera diferente por el hecho de esperar o no esperar determinadas respuestas de determinados grupos. Por ejemplo, si un investigador está estudiando los efectos secundarios de un nuevo medicamento y sabe que estás en el grupo experimental, quizá espere que te encuentres mal y por ese motivo te preste más atención que si supiera que estás en el grupo de control. Esto puede dar lugar a datos sesgados y resultados engañosos.

Si el investigador sabe qué tratamiento está recibiendo cada persona pero los sujetos no lo saben, se dice que el estudio está *enmascarado* (no doblemente enmascarado; o ciego simple). Los estudios enmascarados son mejor que nada, pero lo ideal son los estudios doblemente enmascarados. A lo mejor se te ocurre la siguiente pregunta: en un estudio con doble enmascaramiento, ¿hay alguien que sí sepa qué tratamiento está recibiendo cada sujeto? La respuesta es sí. Lo normal es que haya un tercero, por ejemplo un ayudante de laboratorio, que tenga esa información.

En algunos casos los sujetos saben en qué grupo están porque no hay manera de ocultárselo; por ejemplo si se comparan los beneficios de practicar yoga contra los de practicar *jogging*. Sin embargo, el sesgo es menor cuando no se comunica a los sujetos el propósito exacto del estudio. Esta forma irregular de proceder debe ser supervisada por una junta de revisión institucional para tener la certeza de que no va en contra de la ética. (Relee el apartado anterior "Problemas éticos".)

Analizar los datos correctamente

Una vez recopilados, los datos se meten en esa misteriosa caja llamada *análisis estadístico*, donde se realizan los cálculos numéricos. La elección de un análisis es igual de importante (en lo que respecta a la calidad de los resultados) que cualquier otro aspecto de un estudio. El análisis debe planificarse con antelación, durante la fase de diseño del experimento. De este modo se evitan problemas graves después de haber recopilado los datos. Cuando vayas a seleccionar la forma de análisis, hazte la siguiente pregunta: "Después de haber analizado los datos, ¿podré responder de manera legítima y correcta a la pregunta que he formulado?". Si la respuesta es "no", significa que ese análisis no es el adecuado.

Entre los tipos de análisis básicos están los *intervalos de confianza* (se utilizan para estimar un valor poblacional, o la diferencia entre dos valores poblacionales); los *contrastes de hipótesis* (se utilizan para contrastar una afirmación referida a una o dos poblaciones, por ejemplo, que un medicamento es más eficaz que otro); y los *análisis de correlación y regresión* (se utilizan para determinar si una variable cuantitativa puede predecir o causar cambios en otra variable cuantitativa o de qué forma lo hace). En los capítulos 13, 15 y 18, respectivamente, encontrarás información sobre estos tipos de análisis.

Cuando estés decidiendo la forma en que vas a analizar los datos, debes asegurarte de que los datos y el tipo de análisis son compatibles. Por ejemplo, si quieres comparar un grupo experimental con un grupo de control para saber cuánto han adelgazado los sujetos tras seguir una nueva dieta (frente a una dieta ya existente), tendrás que recopilar datos sobre los kilos que adelgazó cada persona, no simplemente el peso de la persona al término del estudio.

Extraer conclusiones apropiadas

En mi opinión, los principales errores que cometen los investigadores al extraer conclusiones sobre sus estudios son los siguientes (los comento en los apartados siguientes):

- ✔ Exagerar los resultados.
- ✔ Establecer relaciones o dar explicaciones que no están respaldadas por las estadísticas.
- ✔ Ir más allá del ámbito del estudio en cuanto a la aplicabilidad de los resultados.

Exagerar los resultados

Muchas veces los medios de comunicación difunden noticias que exageran los resultados de una investigación. Cuando leas una noticia o te hablen de un estudio, entérate de cómo lo hicieron y a qué conclusiones llegaron exactamente.

Los comunicados de prensa también tienden a exagerar los resultados. Por ejemplo, en un reciente comunicado de prensa del Instituto Estadounidense sobre el Abuso de Drogas (NIDA), los investigadores afirmaban que el consumo de éxtasis había disminuido con respecto al año anterior. Sin embargo, según pude ver en los resultados estadísticos del informe, aunque el porcentaje de adolescentes de la muestra que dijeron consumir éxtasis fue inferior al del año anterior, los investigadores consideraron que esa diferencia no era estadísticamente significativa cuando intentaron extrapolarla a la población formada por todos los adolescentes del país. Es decir: aun siendo cierto que en la muestra de ese año hubo menos adolescentes que consumieron éxtasis, la diferencia no puede atribuirse más que a la variabilidad de una muestra a otra. (En el capítulo 14 encontrarás más información sobre la significación estadística.)

Los títulos y los párrafos iniciales de los comunicados de prensa y los artículos de periódico generalmente exageran los resultados de los estudios. Los hallazgos espectaculares y los grandes avances científicos proporcionan muy buenos titulares, y muchas veces los periodistas sacan noticias de donde no las hay. ¿Cómo distinguir entre la verdad y la exageración? Lo mejor es leer la letra pequeña.

Llevar los resultados un paso más allá de los datos reales

El estudio que relaciona la maternidad tardía con una mayor longevidad ilustra otro error común de los investigadores. ¿Los resultados de ese estudio de observación permiten inferir que dar a luz a una edad tardía puede hacer que las mujeres vivan más años? Los investigadores dijeron que no. Su explicación de los resultados fue que tener un hijo a una edad más avanzada de lo habitual pudo deberse a que esas mujeres tuvieran un reloj biológico "más lento", lo cual podría ralentizar el proceso de envejecimiento.

Mi pregunta para esos investigadores es la siguiente: "Entonces, ¿por qué no estudiasteis eso en lugar de fijaros tan sólo en la edad?". El estudio no incluía ninguna información que pudiera llevarme a concluir que las mujeres que tuvieron un hijo después de cumplir los cuarenta envejecieron más despacio que otras mujeres. Por tanto, a mi juicio, los investigadores no deberían haber llegado a esa conclusión. O bien deberían haber indicado claramente que se trata de una simple teoría que requiere más estudios. Si nos basamos en los datos de ese estudio, la teoría de los investigadores es un esfuerzo de fe (de todos modos, como yo tuve a mi hijo a los cuarenta y uno, me gustaría que tuvieran razón).

Es habitual que en un comunicado de prensa o en un artículo de periódico el investigador explique por qué razón cree que los resultados del estudio fueron los que fueron y qué implicaciones tienen esos resultados para la sociedad en su conjunto, por mucho que no haya estudiado aún esa razón. Estas explicaciones pueden darse en respuesta a las preguntas del periodista sobre la investigación, preguntas que luego se eliminan

para dejar tan sólo las declaraciones más jugosas del investigador. Muchas de estas explicaciones a posteriori no son más que teorías todavía por verificar. En estos casos, desconfía de las conclusiones, explicaciones o asociaciones que no estén respaldadas por los estudios.

Los periodistas quieren que leas el artículo (para eso les pagan), de manera que utilizan titulares sensacionalistas o dan como ciertas relaciones "causaefecto" no confirmadas, porque su trabajo consiste en venderte la noticia. Tu trabajo, en cambio, consiste en desconfiar.

Generalizar los resultados a personas que escapan al alcance del estudio

Únicamente puedes extraer conclusiones sobre la población que está representada por la muestra. Si en tu muestra sólo hay hombres, no puedes extraer conclusiones sobre mujeres. Si en tu muestra sólo hay personas jóvenes y sanas, no puedes extraer conclusiones sobre todo el mundo. Sin embargo, eso es justo lo que intentan hacer muchos investigadores, que pueden acabar obteniendo resultados engañosos.

Para determinar si las conclusiones de un investigador son válidas (en el capítulo 16 encontrarás más información sobre muestras y poblaciones), haz lo siguiente:

1. **Averigua cuál es la población objetivo (es decir, el grupo sobre el cual el investigador quiere extraer conclusiones).**
2. **Averigua cómo se seleccionó la muestra y comprueba si dicha muestra es representativa de la población objetivo (y no de una población más pequeña).**
3. **Verifica las conclusiones a las que han llegado los investigadores y asegúrate de que no están intentando extrapolar sus resultados a una población más grande de la que realmente han estudiado.**

Tomar decisiones con conocimiento de causa

Que alguien diga que ha hecho un "estudio científico" o un "experimento científico" no es razón para pensar que lo ha hecho bien o que los resultados son fiables (tampoco estoy diciendo que debas rechazar todo lo que veas y oigas). Por desgracia, cuando era asesora en asuntos relacionados con la estadística me encontré con muchos experimentos mal hechos. Lo peor de todo es que si un experimento no se ha hecho bien, ya no tiene remedio. Sólo te queda no hacer caso de los resultados.

Aquí tienes algunos consejos para decidir con conocimiento de causa si puedes o no creer los resultados de un experimento, sobre todo si esos resultados son muy importantes para ti:

- ✔ **Cuando oigas o veas el resultado por primera vez, coge un lápiz y apunta lo que hayas oído o visto, dónde lo oíste o viste, quién hizo la investigación y cuáles fueron los principales resultados** (por esta razón yo siempre tengo papel y lápiz en el cuarto de estar y en el bolso).
- ✔ **Encuentra a la persona que hizo la investigación original y pídele una copia del informe o artículo.**
- ✔ **Lee el informe y comprueba si el experimento cumple los ocho criterios descritos en el apartado "Diseñar un experimento válido" de este mismo capítulo** (para hacer esto no es necesario que entiendas todo lo que pone en el informe).
- ✔ **Examina detenidamente las conclusiones que extrae el investigador a partir de sus observaciones.** Muchos investigadores tienden a exagerar los resultados o extraer conclusiones no respaldadas por datos estadísticos, o bien intentan extrapolar sus resultados a una población mayor que la estudiada.
- ✔ **No tengas miedo de hacer preguntas a los medios, a los investigadores o incluso a los expertos que conozcas.** Por ejemplo, si tienes una pregunta sobre un estudio clínico, pregúntale a tu médico. ¡Seguro que estará encantado de tener un paciente tan bien informado!
- ✔ **Y por último, no te pases de escéptico sólo porque ahora eres más consciente de que mucha gente hace las cosas mal.** No todo está mal. Hay muchos más investigadores buenos, resultados fiables y periodistas bien informados que lo contrario. Debes ser prudente y estar preparado para detectar posibles problemas, pero sin rechazarlo todo de buenas a primeras.

Capítulo 18

Buscar vínculos: correlación y regresión

En este capítulo

- Explorar las relaciones estadísticas entre variables numéricas
- Estudiar la correlación y la regresión lineal
- Realizar predicciones basadas en relaciones conocidas
- Comprender la diferencia entre correlación y causalidad

Los medios de comunicación nos informan puntualmente sobre todas las relaciones observadas por los investigadores. Hoy mismo he oído que el uso frecuente de videojuegos puede afectar negativamente a la capacidad de concentración de los niños, que la cantidad de cierta hormona en el organismo femenino puede predecir el momento de aparición de la menopausia, y que cuanto más deprimido estás, más chocolate comes, y cuanto más chocolate comes, más te deprimes (¡qué deprimente!).

Algunos estudios son legítimos y ayudan a mejorar nuestra calidad de vida y nuestra longevidad. Otros, en cambio, no son nada fiables. Por ejemplo, un estudio dice que hacer veinte minutos de ejercicio físico tres veces por semana es mejor que sesenta minutos una vez por semana, otro estudio dice lo contrario, y un tercer estudio dice que no existe diferencia.

Si todos estos resultados contradictorios te tienen confundido, no te preocupes. En este capítulo aprenderás a analizar y evaluar las afirmaciones de los investigadores y a tomar tus propias decisiones sobre los titulares y las noticias que te presentan correlaciones a diario. Descubrirás qué significa realmente que dos variables estén correlacionadas, cuándo puede concluirse que existe una relación de causa y efecto, y cuándo y cómo predecir una variable a partir de otra.

Representar una relación con un diagrama de dispersión

El otro día me llamó la atención un artículo de la revista *Garden Gate*: "Cómo saber la temperatura contando los chirridos de un grillo". Según el artículo, sólo tienes que encontrar un grillo, contar el número de veces que chirría en 15 segundos, sumar 40 y ya tienes una estimación de la temperatura en grados Fahrenheit. Si quieres convertir los grados Fahrenheit a centígrados, utiliza esta sencilla fórmula: (Temp. en °F 32) x 5/9.

El Servicio Estadounidense de Meteorología tiene incluso un "convertidor de chirridos" en su sitio web. Introduces el número de chirridos en 15 segundos y el convertidor te da la temperatura estimada en cuatro unidades diferentes, incluidos grados Fahrenheit y grados centígrados.

Se han hecho varios estudios que confirman esta relación entre la frecuencia de los chirridos y la temperatura. A modo de ejemplo, he tomado tan sólo un subconjunto de los datos disponibles (mira la tabla 181).

Tabla 18-1. Datos sobre chirridos y temperaturas (extracto)

Número de chirridos (en 15 segundos)	*Temperatura (grados Fahrenheit)*
18	57
20	60
21	64
23	65
27	68
30	71
34	74
39	77

Cada observación consta de dos variables que están relacionadas: el número de veces que el grillo chirría en 15 segundos (la variable *X*) y la temperatura en el momento de recopilarse los datos (la variable *Y*). En estadística, este tipo de datos bidimensionales se denominan *datos bivariables*. Cada observación contiene una pareja de datos recopilados de forma simultánea. Por ejemplo, la fila uno de la tabla 181 muestra una pareja de datos (18, 57).

Los datos bivariables generalmente se organizan en un gráfico que los estadísticos llaman *diagrama de dispersión*. Un diagrama de dispersión tiene dos dimensiones, una dimensión horizontal (el eje *X*) y una dimensión vertical (el eje *Y*). Ambos ejes son numéricos: cada uno contiene una línea numerada. En los siguientes apartados explico la forma de elaborar e interpretar un diagrama de dispersión.

Elaborar un diagrama de dispersión

Ubicar observaciones (o puntos) en un diagrama de dispersión es parecido al juego de hundir la flota. Cada observación tiene dos coordenadas; la primera corresponde al primer dato de la pareja (la coordenada *X*; cuánto te desplazas a la izquierda o a la derecha). La segunda coordenada corresponde al segundo dato de la pareja (la coordenada *Y*; cuánto te desplazas hacia arriba o hacia abajo). El punto que representa a esa observación se ubica en la intersección de las dos coordenadas.

La figura 181 muestra un diagrama de dispersión para los datos sobre chirridos y temperaturas incluidos en la tabla 181. Como he ordenado los datos en función de sus valores *X*, los puntos del diagrama se corresponden, de izquierda a derecha, con las observaciones que aparecen en la tabla 181, en ese mismo orden.

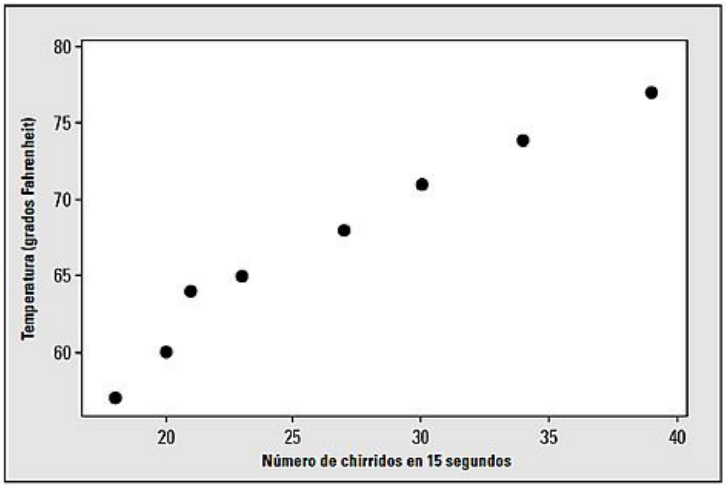


Figura 18-1: Diagrama de dispersión de los chirridos de un grillo en relación con la temperatura ambiente

Interpretar un diagrama de dispersión

Para interpretar un diagrama de dispersión hay que buscar tendencias en los datos de izquierda a derecha:

- ✔ Si los datos muestran una tendencia ascendente de izquierda a derecha, significa que existe una *relación positiva entre X e Y*. A media que los valores *X* aumentan (se desplazan a la derecha), los valores *Y* también aumentan (se desplazan hacia arriba) en cierta medida.
- ✔ Si los datos muestran una tendencia descendente de izquierda a derecha, significa que existe una *relación negativa entre X e Y*. A media que los valores *X* aumentan (se desplazan a la derecha), los valores *Y* disminuyen (se desplazan hacia abajo) en cierta medida.
- ✔ Si los datos no parecen revelar ninguna tendencia (por vaga que sea), entonces no existe ninguna relación entre *X* e *Y*.

Una tendencia de particular interés es la tendencia *lineal*, cuando los datos parecen dibujar una línea ascendente o descendente. En la figura 181 se aprecia una relación lineal positiva entre el número de chirridos y la temperatura. Es decir, si el número de chirridos aumenta, la temperatura también aumenta.

En este capítulo me refiero exclusivamente a las relaciones lineales. Se dice que existe una *relación lineal entre X e Y* cuanto los valores *X* y los valores *Y* parecen dibujar una línea, ya sea ascendente (con pendiente positiva) o descendente (con pendiente negativa). Además de las tendencias lineales ascendentes/descendentes, hay otros tipos de tendencias (por ejemplo curvas o funciones exponenciales); sin embargo, escapan al alcance de este libro. Lo bueno es que el caso de las líneas ascendentes/ descendentes incluye muchos tipos de relaciones.

Los diagramas de dispersión muestran posibles asociaciones o relaciones entre dos variables. No obstante, el hecho de que un gráfico o diagrama revele que algo está pasando no significa que exista una relación de causa y efecto.

Por ejemplo, un médico observa que las personas que toman vitamina C todos los días parecen resfriarse menos. ¿Significa esto que la vitamina C previene los resfriados? No necesariamente. Puede que las personas que se preocupan más por su salud tomen vitamina C a diario, pero también cuidan la alimentación, no tienen sobrepeso, hacen ejercicio todos los días y se lavan las manos más a menudo. Si el médico quiere

saber si la causa es la vitamina C, tendrá que realizar un experimento bien diseñado que descarte esos otros factores. (Encontrarás más información al respecto en el apartado "Explicar la relación: correlación frente a causa y efecto", más adelante.)

Cuantificar relaciones lineales por medio de la correlación

Cuando tienes los datos bivariables representados gráficamente en un diagrama de dispersión (relee el apartado anterior), si observas algún tipo de tendencia lineal, el siguiente paso es obtener algunas estadísticas que cuantifiquen o midan el alcance y la naturaleza de la relación. En los siguientes apartados hablo de la *correlación*, un estadístico que mide la fuerza y la dirección de una relación lineal entre dos variables. En particular, explico la forma de calcular e interpretar la correlación y cuáles son sus propiedades más importantes.

Calcular la correlación

En el apartado anterior, "Interpretar un diagrama de dispersión", decía que los datos que se parecen a una línea ascendente presentan una relación lineal positiva, y los que se parecen a una línea descendente presentan una relación lineal negativa. Sin embargo, no explicaba cuándo la relación lineal era fuerte o débil. La fuerza de una relación lineal depende de cuánto se parezcan los datos a una línea, y, naturalmente, existen varios niveles de similitud.

¿Puede un solo estadístico medir la fuerza y la dirección de una relación lineal entre dos variables? La respuesta es sí. Los estadísticos utilizan el *coeficiente de correlación* para medir la fuerza y la dirección de la relación lineal entre dos variables numéricas *X* e *Y*. El coeficiente de correlación para una muestra de datos se designa con la letra *r*.

Aunque la definición común de *correlación* se aplica a dos elementos cualesquiera que estén relacionados (por ejemplo, el sexo y la filiación política), los estadísticos utilizan este término exclusivamente para expresar la relación entre dos variables numéricas. El término formal para indicar la intensidad de esa relación es *coeficiente de correlación*. Se han creado muchas medidas de correlación diferentes; la utilizada en este caso se llama *coeficiente de correlación de Pearson* (pero en adelante me referiré a él simplemente como correlación).

La fórmula para la correlación (*r*) es

$$r=\frac{1}{n-1}\left(\frac{\sum_x\sum_y(x-\bar{x})(y-\bar{y})}{s_x s_y}\right)$$

en la que *n* es el número de parejas de datos; $\bar{x}$ y $\bar{y}$ son las medias muestrales de todos los valores *x* y todos los valores *y*, respectivamente; y s_x y s_y son las desviaciones estándares de todos los valores *x* y de todos los valores *y*, respectivamente.

Para calcular la correlación, *r*, a partir de un conjunto de datos, sigue los pasos siguientes:

1. **Encuentra la media de todos los valores *x* ($\bar{x}$) y la media de todos los valores *y* ($\bar{y}$).**

   En el capítulo 5 encontrarás más información sobre cómo calcular la media.

2. **Encuentra la desviación estándar de todos los valores *x* (llámala s_x) y la desviación estándar de todos los valores *y* (llámala s_y).**

   En el capítulo 5 explico la forma de calcular la desviación estándar.

3. **Para cada pareja *(x, y)* del conjunto de datos, calcula *x* menos $\bar{x}$ e *y* menos $\bar{x}$ y luego multiplica ambos números, para obtener $(x-\bar{x})(y-\bar{y})$.**

4. **Suma todos los resultados del paso 3.**

5. **Divide el resultado de la suma entre $s_x \times s_y$.**

6. **Divide ese resultado entre *n*–1, donde *n* es el número de parejas *(x, y)* (es lo mismo que multiplicar por 1 partido por *n*–1).**

   El resultado es la correlación, *r*.

Por ejemplo, supongamos que tienes el conjunto de datos (3, 2), (3, 3) y (6, 4). El coeficiente de correlación *r* se calcula de la manera siguiente (ten en cuenta que, para estos datos, los valores *x* son 3, 3, 6, y los valores *y* son 2, 3, 4):

1. **$\bar{x}$ es 12 / 3 = 4, y $\bar{y}$ es 9 / 3 = 3.**

2. **Las desviaciones estándares son s_x = 1,73 y s_y = 1,00.**

Los cálculos se explican paso a paso en el capítulo 5.

3. **Si multiplicas las diferencias calculadas en el paso 3, tienes: (3 – 4) (2 – 3) = (– 1) (– 1) = +1; (3 – 4) (3 – 3) = (– 1) (0) = 0; (6 – 4) (4 – 3) = (2) (1) = + 2.**

4. **Si sumas los resultados del paso 3, resulta 1 + 0 + 2 = 3.**

5. **Al dividir entre $s_x \times s_y$, resulta 3 / (1,73 × 1) = 3 / 1,73 = 1,73.**

6. **Ahora divide el resultado del paso 5 entre 3 – 1 (que es 2), y obtienes la correlación *r* = 0,87.**

Interpretar la correlación

La correlación *r* siempre está comprendida entre +1 y –1. Para interpretar varios valores de *r* (aquí no existen regla fijas, tan sólo la regla práctica de Rumsey), tienes que ver cuál de los siguientes valores es el más próximo a tu correlación:

- ✔ **Exactamente –1.** Una relación lineal descendente (negativa) perfecta.
- ✔ **–0,70.** Una relación lineal descendente (negativa) fuerte.
- ✔ **–0,50.** Una relación lineal descendente (negativa) moderada.
- ✔ **–0,30.** Una relación lineal descendente (negativa) débil.
- ✔ **0.** No existe relación lineal.
- ✔ **+0,30.** Una relación lineal ascendente (positiva) débil.
- ✔ **+0,50.** Una relación lineal ascendente (positiva) moderada.
- ✔ **+0,70.** Una relación lineal ascendente (positiva) fuerte.
- ✔ **Exactamente +1.** Una relación lineal ascendente (positiva) perfecta.

Si el diagrama de dispersión no indica la existencia de una relación lineal, por débil que sea, la correlación no tiene apenas significado. ¿Por qué medir la fuerza de la relación lineal si de entrada esa relación es casi inexistente? No obstante, la no existencia de una relación lineal puede verse de dos maneras: 1) si no existe ninguna relación en absoluto, calcular la correlación no tiene sentido porque la correlación únicamente se aplica a las relaciones lineales; y 2) si existe una relación fuerte pero no es lineal, la correlación puede inducir a error, ya que en algunos casos existe una relación curva fuerte y la correlación también es fuerte. Por eso es imprescindible examinar primero el diagrama de dispersión.

La figura 182 muestra los diagramas de dispersión correspondientes a varias correlaciones, para que puedas ver la fuerza y la dirección de la relación. La figura 182a muestra una correlación de +1, la figura 182b muestra una correlación de –0,50, la figura 182c muestra una correlación de +0,85, y la figura 182d muestra una correlación de +0,15. Si te fijas en las figuras 182a y c, verás que la figura 182a es una línea recta ascendente perfecta, y la figura 182c muestra una tendencia lineal ascendente muy fuerte. La figura 182b es descendente, pero los puntos están dispersos en una franja más ancha, lo cual indica la existencia de una relación lineal, pero no tan fuerte como en las figuras 182a y 182c. En la figura 182d no se aprecia apenas nada (es lógico, puesto que su correlación está muy próxima a 0).

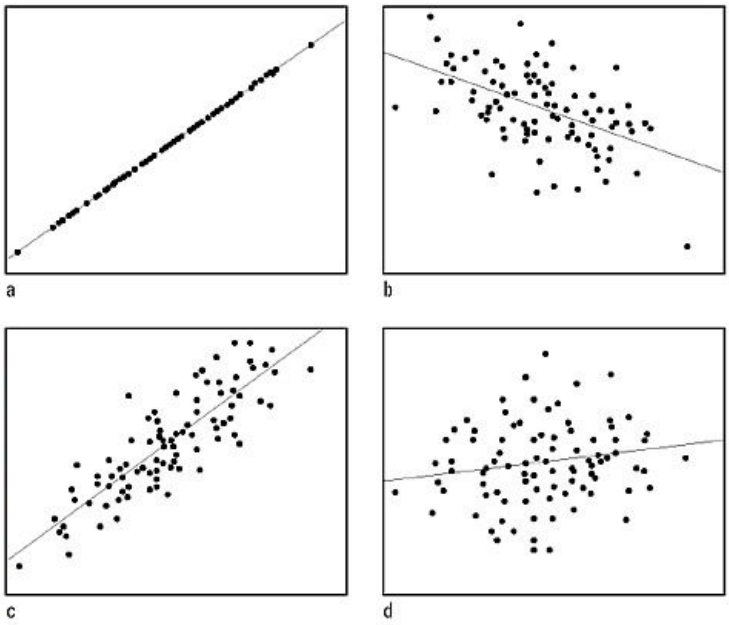


Figura 18-2: Diagramas de dispersión con correlaciones de a) +1; b) –0,5; c) +0,85; y d) +0,15

Mucha gente comete el error de creer que una correlación de –1 es algo malo, pues indica la ausencia de relación. En realidad es justo al revés. Una correlación de –1 significa que los datos forman una línea recta perfecta, la relación lineal más fuerte que puede existir. El signo "–" (menos) simplemente indica que la relación es negativa, una línea descendente.

¿Cuándo se considera que el valor está lo suficientemente cerca de –1 o de +1 como para indicar una relación lineal suficientemente fuerte? La mayoría de los estadísticos no se emocionan demasiado hasta ver una correlación de al menos +0,5 o –0,5. De todos modos, no esperes que una correlación llegue siempre a 0,99. Recuerda que manejas datos reales, y los datos reales no son perfectos.

Para mi subconjunto de datos sobre chirridos de grillo y temperaturas, en el apartado anterior "Representar una relación con un diagrama de dispersión", calculé una correlación de 0,98, que casi nunca se da en el mundo real (¡bien por los grillos!).

Examinar las propiedades de la correlación

El coeficiente de correlación tiene varias propiedades importantes:

- ✔ La correlación siempre está comprendida entre –1 y +1, como te he explicado en el apartado anterior.
- ✔ La correlación es una medida adimensional, lo cual significa que si cambias las unidades de *X* o de *Y*, la correlación no varía. Por ejemplo, cambiar la temperatura de grados Fahrenheit a grados centígrados no afecta a la correlación existente entre la frecuencia de los chirridos *(X)* y la temperatura exterior *(Y)*.
- ✔ Las variables *X* e *Y* pueden intercambiarse en el conjunto de datos sin que la correlación varíe. Por ejemplo, si la altura y el peso tienen una correlación de 0,53, el peso y la altura tienen la misma correlación.

Trabajar con la regresión lineal

En el caso de dos variables numéricas *X* e *Y*, cuando se ha establecido una correlación al menos moderada a través de la propia correlación y el diagrama de dispersión, sabes que dichas variables presentan algún tipo de relación lineal. Los investigadores se sirven a menudo de esa relación para predecir el valor (medio) de *Y* a partir de un valor *X* conocido, utilizando una línea recta. Esta línea se llama *recta de regresión*. Si conoces la pendiente y el punto de intersección con el eje *y* (ordenada en el origen) de esa recta de regresión, entonces puedes introducir un valor de *X* y obtienes el valor medio de *Y*. Dicho de otro modo, predices el valor (medio) de *Y* a partir del valor de *X*. En los siguientes apartados ofrezco unas nociones básicas sobre la ecuación de regresión lineal (más adelante en este mismo capítulo explico cómo realizar predicciones con la regresión lineal).

No hagas un análisis de regresión a menos que previamente hayas encontrado una correlación al menos moderada entre las dos variables (como norma general, yo siempre establezco que esta correlación debe ser de al menos +0,50 o –0,50, pero otros estadísticos pueden aplicar criterios diferentes). ¡He visto casos en que los investigadores hacen predicciones con una correlación de tan sólo 0,20! Lo mires como lo mires, eso es una barbaridad. Si la nube de datos ya ni siquiera se parece a una línea, no puedes utilizar una línea para hacerlos encajar y realizar predicciones (y aun así, hay que gente que lo intenta).

Determinar qué variable es *X* y cuál es *Y*

Antes de pasar a la ecuación para la recta de regresión, tienes que identificar cuál de las dos variables es *X* y cuál es *Y*. Cuando analizas correlaciones (como explico anteriormente en este capítulo), no importa a qué variable llames *X* y a cuál llames *Y*, siempre y cuando mantengas la coherencia para todos los datos. Sin embargo, cuando tratas de encajar líneas y hacer predicciones, la elección de *X* y de *Y* sí que importa.

Entonces, ¿cómo determinas cuál es cuál? En general, *Y* es la variable que quieres predecir, y *X* es la variable que utilizas para realizar esa predicción. En el ejemplo anterior de los grillos, utilizas el número de chirridos para predecir la temperatura. Por tanto, en este caso la variable *Y* es la temperatura, y la variable *X* es el número de chirridos. Como consecuencia, es posible predecir *Y* a partir de *X* utilizando la ecuación de una recta siempre y cuando exista una relación lineal suficientemente fuerte.

En estadística, la variable *X* (los chirridos, en mi ejemplo anterior) se llama la *variable explicativa*, porque si *X* cambia, la pendiente indica (o explica) cuánto se espera que cambie *Y* en consecuencia. Así pues, siguiendo esta misma lógica, la variable *Y* se llama la *variable explicada*. Otros nombres para *X* e *Y* son la *variable independiente* y la *variable dependiente*, respectivamente.

Comprobar las condiciones

En el caso de dos variables numéricas, puedes encontrar una línea que te permita predecir *Y* a partir de *X*, siempre y cuando se cumplan las dos condiciones siguientes (extraídas de los apartados anteriores):

- ✔ El diagrama de dispersión presenta una tendencia lineal.
- ✔ La correlación, *r*, es entre moderada y fuerte (por regla general, superior a 0,50 o –0,50).

Algunos investigadores no comprueban estas condiciones antes de hacer predicciones. Sin embargo, dichas predicciones no serán válidas a menos que las dos condiciones se cumplan.

Pero supongamos que la correlación es fuerte: ¿igualmente hay que fijarse en el diagrama de dispersión? Sí. Hay situaciones en que los datos presentan una forma curvada y la correlación continúa siendo fuerte. En esos casos tampoco es válido hacer predicciones utilizando una línea recta. Las predicciones deben realizarse a partir de una curva. (Este tema escapa al alcance de este libro; si estás interesado, te recomiendo *Statistics II for Dummies*, donde hablo de las relaciones no lineales.)

Calcular la recta de regresión

El diagrama de dispersión de la figura 181 revela una clara tendencia lineal. Como vimos anteriormente en este mismo capítulo, la correlación entre chirridos y temperatura es muy fuerte (r=0,98). Sabiendo esto, ahora puedes encontrar la línea recta que mejor se ajusta a los datos (la que se encuentra a menos distancia del conjunto de puntos en general). En estadística, esta técnica de encontrar la recta que mejor se ajusta se denomina *análisis de regresión lineal simple utilizando el método de los mínimos cuadrados*.

La fórmula para la *recta de mejor ajuste* (o *recta de regresión*) es $y=mx+b$, en la que m es la pendiente de la recta y b es el punto de intersección con el eje y (ordenada en el origen). Aunque esta ecuación es la misma que se utiliza en álgebra para encontrar una línea recta, debes tener presente que en estadística los puntos no dibujan una línea perfecta; la línea es un modelo alrededor del cual están situados los datos cuando existe una tendencia lineal fuerte.

- ✔ La *pendiente* de una línea recta es el cociente entre la variación de *Y* y la variación de *X*. Por ejemplo, una pendiente de 10 / 3 significa que si el valor *x* aumenta (se desplaza hacia la derecha) 3 unidades, el valor *y* se desplaza hacia arriba 10 unidades en promedio.
- ✔ La intersección con el eje *y*, también llamada ordenada en el origen, es el punto del eje *y* donde el valor de *x* es cero. Por ejemplo, en la ecuación 2*x*–6, la línea corta el eje *y* en el punto –6. Las coordenadas de este punto son (0, –6). Por tanto, cuando un línea corta el eje *y*, el valor *x* siempre es 0.

Para encontrar la recta de mejor ajuste, tienes que encontrar los valores de *m* y *b* que mejor se ajusten a la tendencia de los datos, de acuerdo con tus criterios. Existen diferentes criterios que pueden conducir a otras rectas, pero en este libro (y en todos los cursos de introducción a la estadística, en general) utilizo el criterio de encontrar la recta que reduzca al mínimo lo que en estadística se llama la *suma de cuadrados residual (SCR)*. La SCR es la suma de los cuadrados de todas las diferencias entre los puntos de la recta propuesta y los puntos reales del conjunto de datos. La recta que tenga la SCR más baja posible es la vencedora, y su ecuación se utiliza como recta de mejor ajuste. Este proceso es el que da nombre al *método de mínimos cuadrados*.

A lo mejor estás pensando que hace falta probar montones de rectas diferentes para ver cuál se ajusta mejor. Por suerte existe una solución más sencilla (aunque es cierto que visualizar una línea recta en el diagrama de dispersión ayuda a pensar en cuál debería ser la respuesta). La recta de mejor ajuste tiene una pendiente y una ordenada en el origen concretas que pueden calcularse mediante fórmulas (los cálculos no son demasiado difíciles).

Para ahorrar tiempo calculando la recta de mejor ajuste, encuentra primero los cinco estadísticos que necesitarás para tus cálculos:

1. La media de los valores *x* (designada como $\bar{x}$).
2. La media de los valores *y* (designada como $\bar{y}$).
3. La desviación estándar de los valores *x* (designada como s_x).
4. La desviación estándar de los valores *y* (designada como s_y).
5. La correlación entre *X* e *Y* (designada como *r*).

Encontrar la pendiente

La fórmula para la pendiente, *m*, de la recta de mejor ajuste es

$$m = r\left(\frac{s_y}{s_x}\right)$$

en la que *r* es la correlación entre *X* e *Y*, y s_x y s_y son las desviaciones estándares de los valores *x* y de los valores *y*, respectivamente. Sólo hay que dividir s_y entre s_x y multiplicar el resultado por *r*.

Ten en cuenta que la pendiente de la recta de mejor ajuste puede ser un número negativo, ya que la correlación puede ser un número negativo. Una pendiente negativa indica que la recta es descendente. Por ejemplo, existe una relación lineal entre el incremento de agentes de policía y la disminución del número de delitos; en este caso la correlación y, por tanto, la pendiente de la recta de mejor ajuste son negativas.

La correlación y la pendiente de la recta de mejor ajuste son cosas distintas. La fórmula para la pendiente coge la correlación (una medida adimensional) y le asigna unidades. Para entenderlo mejor, considera que s_y/s_x es la variación de *Y* partido por la variación de *X*, en unidades de *X* y de *Y*. Por ejemplo, la variación de la temperatura (en grados centígrados) partido por la variación del número de chirridos (en 15 segundos).

Encontrar la ordenada en el origen

La fórmula para la ordenada en el origen, *b*, de la recta de mejor ajuste es $\boldsymbol{b = \bar{y} - m\bar{x}}$, donde $\bar{x}$ y $\bar{y}$ son las medias de los valores *x* y de los valores *y*, respectivamente, y *m* es la pendiente (cuya fórmula encontrarás en el apartado anterior).

Así pues, para calcular la ordenada en el origen, *b*, de la recta de mejor ajuste hay que empezar por encontrar la pendiente, *m*, de la recta de mejor ajuste siguiendo los pasos indicados en el apartado anterior. Luego hay que multiplicar *m* por $\bar{x}$ y restar el resultado de $\bar{y}$.

Calcula siempre la pendiente antes que la ordenada en el origen. ¡La fórmula para la ordenada en el origen contiene la pendiente!

Interpretar la recta de regresión

Más importante aún que saber calcular la pendiente y la ordenada en el origen para formar la recta de regresión, es saber interpretar sus valores. En los siguientes apartados me refiero a esta cuestión.

Interpretar la pendiente

Algebraicamente, la pendiente entre dos puntos se interpreta como la *distancia vertical dividida por la distancia horizontal*. Si, por ejemplo, la pendiente es 2, puedes expresarlo como 2/1 y decir que, al moverte de un punto a otro de la línea, cuando el valor de la variable *X* aumenta una unidad, el valor de la variable *Y* aumenta dos unidades. En el contexto de una regresión, la pendiente es la parte fundamental de la ecuación porque te dice cuánto cabe esperar que varíe *Y* al aumentar *X*.

En general, la pendiente se mide en unidades de la variable *Y* por unidades de la variable *X*. Se trata de la proporción entre la variación de *Y* y la variación de *X*. Supongamos que un investigador está estudiando el efecto que tienen diferentes dosis en miligramos (mg) de un fármaco sobre la tensión arterial sistólica (se mide en milímetros de mercurio [mmHg]), y observa que la pendiente de la recta de regresión es –2,5. Podemos expresarlo como [–2,5 / 1], y decir que la tensión arterial sistólica previsiblemente disminuirá 2,5 milímetros de mercurio en promedio por cada miligramo que se incremente la dosis administrada del fármaco.

Cuando interpretes la pendiente, asegúrate de utilizar siempre las unidades adecuadas. Si no tienes en cuenta las unidades, no verás la conexión existente entre las dos variables en cuestión. Por ejemplo, si *Y* es la nota de examen y *X* es el tiempo de estudio, y compruebas que la pendiente de la ecuación es 5, ¿qué significa eso? Para saberlo con certeza necesitas conocer las unidades. Una vez te dan esa información, ves que la nota sube cinco puntos (variación de *Y*) por cada hora más de estudio (variación de *X*). Asimismo, debes tener cuidado con las variables que puedan expresarse en más de una unidad (por ejemplo la temperatura, que puede estar en grados Fahrenheit o en grados centígrados). En ese caso, hay que saber qué unidad se está utilizando.

Si utilizar un 1 en el denominador de la pendiente no te resulta muy significativo, puedes multiplicar el numerador y el denominador por cualquier número (siempre que sea el mismo para los dos) y hacer la interpretación correspondiente. En el ejemplo de la tensión arterial sistólica, en lugar de expresar la pendiente como [–2,5 / 1] e interpretarla como una caída de 2,5 milímetros de mercurio por cada incremento de 1 miligramo en la dosis del fármaco, puedes multiplicar el numerador y el denominador por 10 para obtener [–25 / 10], con lo que puedes decir que un incremento de 10 miligramos en la dosis provoca una caída de 25 milímetros de mercurio en la tensión arterial sistólica.

Interpretar la ordenada en el origen

La ordenada en el origen es el punto donde la recta de regresión $y=mx+b$ corta el eje *y* para $x=0$, y se designa con la letra *b*. (Ver el apartado anterior "Encontrar la ordenada en el origen".) A veces la ordenada en el origen puede interpretarse de una manera que tiene sentido, y otras veces no. Esta incertidumbre no existe para la pendiente, que siempre es interpretable. De hecho, entre estos dos elementos, la pendiente es la estrella de la película, mientras que la ordenada en el origen es esa actriz secundaria menos famosa pero también destacable.

Hay ocasiones en que la ordenada en el origen no tiene sentido. Por ejemplo, imagina que utilizas la lluvia para predecir los kilos de maíz que se cosecharán por hectárea de terreno. Sabes que si el conjunto de datos contiene un punto en que la lluvia es 0, los kilos cosechados por hectárea también serán 0. Como resultado, si la recta de regresión corta el eje *y* en algún otro punto, además del 0 (y no hay ninguna garantía de que lo corte en el 0; dependerá de los datos), la ordenada en el origen no tendrá sentido. De forma similar, en esas circunstancias un valor negativo de *y* (producción de maíz) no puede interpretarse.

Otra situación en que no es posible interpretar la ordenada en el origen es cuando no existen datos cerca del punto donde $x=0$. Por ejemplo, supongamos que quieres utilizar las notas obtenidas por los alumnos en el primer trimestre para predecir las notas del segundo trimestre. La ordenada en el origen representa una predicción para el segundo trimestre cuando la nota del primer trimestre es 0. Sin embargo, no cabe esperar que alguien saque un 0 o una nota próxima al 0 en un trimestre a menos que no hiciera el examen, en cuyo caso su nota ya no se habría incluido.

Muchas veces, no obstante, la ordenada en el origen tiene interés y significado, y además has recopilado datos en la zona en la que $x=0$. Por ejemplo, si utilizas la temperatura para intentar predecir cuántos cafés se venderán en los partidos de fútbol que se jueguen en el estadio de Green Bay, Wisconsin (por poner el ejemplo de una localidad donde hace mucho frío), sabes que algunos partidos se jugarán con temperaturas próximas o incluso inferiores a 0 grados centígrados, de manera que predecir las ventas de café a esas temperaturas tiene sentido (como ya

supondrás, cuanto más desciende la temperatura, más cafés se venden).

Un ejemplo ilustrativo: la recta de regresión para los grillos

En el apartado anterior "Representar una relación con un diagrama de dispersión" introduje el ejemplo de los grillos, cuyos chirridos están relacionados con la temperatura. La tabla 18-2 muestra los cinco estadísticos básicos mencionados en "Calcular la recta de regresión", en este caso para el subconjunto de datos sobre grillos. (***Nota:*** redondeo las cifras para facilitar las explicaciones.)

Tabla 18-2. Estadísticos básicos para los datos sobre grillos

Variable	*Media*	*Desviación estándar*	*Correlación*
Número de chirridos	$\bar{x} = 26,5$	$s_x = 7,4$	$r = +0,98$
Temperatura (y)	$\bar{y} = 19,4$	$s_y = 1,97$	

La pendiente, *m*, para la recta de mejor ajuste correspondiente al subconjunto de datos sobre chirridos y temperaturas es $m = r\frac{s_y}{s_x} = 0,98\left(\frac{6,8}{7,4}\right) = 0,9$.

Por tanto, si el número de chirridos aumenta una unidad por cada 15 segundos, cabe esperar que la temperatura aumente 0,26 grados centígrados en promedio. Para que la interpretación tenga más sentido, puedes multiplicar el numerador y el denominador de la pendiente por 10 y entonces decir que si el número de chirridos aumenta 10 unidades (por cada 15 segundos), la temperatura aumentará 2,6 grados centígrados.

Para encontrar la ordenada en el origen, *b*, calculas $\bar{y} - m\bar{x}$; es decir, 19,4– (0,90) (26,5)=43,15. Por consiguiente, la recta de mejor ajuste para predecir la temperatura a partir de los chirridos de grillos es $y=0,90x+43,15$, es decir, la temperatura (en grados centígrados)=0,90*x* (número de chirridos en 15 segundos)+43,2. Llegados a este punto, ¿puedes utilizar la ordenada en el origen para predecir la temperatura cuando no se oyen chirridos de grillo? Como no se han recopilado datos en este punto ni cerca de él, no puedes hacer predicciones para la temperatura en esa zona. Es decir, no puedes predecir la temperatura a partir de los chirridos de los grillos cuando los grillos están callados.

Realizar predicciones adecuadas

Una vez que has determinado la existencia de una relación lineal fuerte y has encontrado la ecuación de la recta de mejor ajuste utilizando $y = mx + b$, utilizas esa recta para predecir el valor *y* (medio) para un valor *x* dado. Para realizar predicciones, introduces el valor *x* en la ecuación y resuelves para *y*. Por ejemplo, si tu ecuación es $y = 2x + 1$ y quieres predecir *y* para $x = 1$, sustituyes la *x* por 1, con lo que te queda $y = 2(1) + 1 = 3$.

Ten en cuenta que eres tú quien elige los valores de *X* (la variable explicativa) que utilizas en la ecuación, y que lo haces para predecir *Y*, la variable explicada, que depende totalmente de *X*. Es decir, utilizas una variable sobre la cual puedes recopilar datos fácilmente con la finalidad de predecir una variable *Y* que es difícil o imposible de medir. Esto funciona bien, siempre y cuando *X* e *Y* estén correlacionadas. En eso precisamente se fundamenta la regresión.

Volviendo a los ejemplos del apartado anterior, la recta de mejor ajuste para los grillos es $y = 0,9x + 43,2$. Imagina que estás de acampada, escuchando los grillos, y recuerdas que es posible predecir la temperatura contando sus chirridos. Cuentas 35 chirridos en 15 segundos, sustituyes la *x* por 35 y te sale que $y = 0,9(35) + 43,2 = 74,7$ (sí, te aprendiste la fórmula de memoria antes de salir de acampada, por si acaso la necesitabas). Por tanto, como los grillos han chirriado 35 veces en 15 segundos, calculas que la temperatura aproximada es de 75 grados Fahrenheit.

Tener una recta de regresión no significa que puedas introducir cualquier valor de *X* y obtener una buena predicción de *Y*. No puedes hacer predicciones utilizando valores *x* que no estén dentro del rango de los valores usados para calcular la regresión. En estadística eso se llama *extrapolación*. Mucho ojo con los investigadores que afirman cosas a partir de datos que no tienen.

Si volvemos al ejemplo de los grillos, vemos que no se han recopilado datos para menos de 18 chirridos ni para más de 39 chirridos en 15 segundos (ver la tabla 18-1). Si intentas hacer predicciones más allá de este intervalo de datos observados, te adentras en terreno desconocido. Cuanto más se alejan de este rango los valores *x*, más dudosas son las predicciones de *y*. ¿Quién puede asegurar que la línea continúa funcionando fuera del intervalo donde se recopilaron los datos? ¿De verdad crees que los grillos pueden chirriar más y más de prisa sin llegar a un límite? ¡Llegará un momento en que se desmayarán o empezarán a arder! ¿Y qué sentido tiene un número de chirridos negativo? (es como preguntarse qué se oye al aplaudir con una sola mano).

No puedes estar seguro de que todos los puntos de datos se ajustarán bien a la recta de regresión, por mucho que la correlación sea alta. Puede que uno o dos puntos se salgan de la tendencia general del resto de los datos; esos puntos se llaman *valores atípicos*. Probablemente uno o dos valores atípicos no influyan demasiado en el ajuste general de la recta de regresión, pero al final puedes ver que la línea no ha dado buen resultado en esos puntos concretos.

La diferencia numérica entre el valor de *y* predicho a partir de la línea y el valor *y* real que has obtenido de tus datos es lo que se llama *residuo*. Los valores atípicos tienen residuos grandes en comparación con el resto de los puntos, de manera que merece la pena investigar si se ha cometido un error en los datos correspondientes a esos puntos o si hay algo de particular interés en dichos datos. (En el libro *Statistics II for Dummies* hablo sobre los residuos con mucho más detenimiento.)

Explicar la relación: correlación frente a causa y efecto

Los diagramas de dispersión y las correlaciones identifican y cuantifican relaciones entre dos variables. Sin embargo, aunque un diagrama de dispersión revele una tendencia clara y se compruebe que los datos presentan una correlación fuerte, eso no significa necesariamente que exista una relación de causa y efecto entre las dos variables. Una *relación de causa y efecto* es aquella en que un cambio en una variable (en este caso *X*) provoca un cambio en otra variable (en este caso *Y*). O dicho de otro modo: el cambio en *Y* no sólo está asociado a un cambio en *X*, sino también causado directamente por *X*.

Por ejemplo, supongamos que se lleva a cabo un experimento médico bien controlado con el fin de determinar los efectos de diferentes dosis de cierto fármaco sobre la tensión arterial. (En el capítulo 17 explico todos los pormenores de los experimentos.) Los investigadores examinan el diagrama de dispersión y ven una tendencia lineal descendente bastante clara; calculan la correlación y resulta ser fuerte. Concluyen entonces que el incremento de la dosis de ese fármaco hace disminuir la tensión arterial. Esta conclusión de causa y efecto es correcta porque en el experimento los investigadores controlaron otras variables que pudieron influir en la tensión arterial, por ejemplo, el consumo de otros fármacos, la edad, el estado de salud general, etc.

Sin embargo, si dibujaras un diagrama de dispersión y examinaras la correlación entre el consumo de helado y la tasa de homicidios en la ciudad de Nueva York, también verías una fuerte relación lineal (en este caso, ascendente). Y aun así, nadie en su sano juicio diría que el mayor consumo de helado causa más homicidios.

¿Qué está ocurriendo aquí? En el primer caso, los datos se recopilaron a través de un experimento médico bien controlado que reduce al mínimo la influencia de otros factores que puedan afectar a la tensión arterial. En el segundo ejemplo, los datos estaban basados exclusivamente en la observación, y no se analizaron otros factores. Posteriormente los investigadores descubrieron la causa de esta fuerte relación: tanto los incrementos de la tasa de homicidios como los incrementos de las ventas de helado están relacionados con incrementos de la temperatura. La temperatura en este caso es una *variable de confusión*; afecta tanto a *X* como a *Y* pero no se incluyó en el estudio. (Si necesitas refrescar este concepto ve al capítulo 17.)

Que se concluya o no la existencia de una relación causal entre dos variables depende de la manera en que se haya llevado a cabo el estudio. He visto a mucha gente proclamar la existencia de una relación de causa y efecto con sólo examinar un diagrama de dispersión o una correlación. ¿Por qué lo hacen? Porque es lo que quieren creer (se aplican el dicho al revés, de manera que en su caso es "creer para ver"). Mucho ojo con esta táctica. A fin de establecer una relación de causa y efecto, necesitas un experimento bien diseñado o bien un montón de estudios de observación. Si alguien intenta establecer una relación de causa y efecto simplemente mostrando un diagrama o gráfico, preocúpate de averiguar cómo se diseñó el estudio y cómo se recopilaron los datos, y luego evalúa el estudio de forma adecuada aplicando los criterios expuestos en el capítulo 17.

Algunos investigadores y periodistas no hacen ningún caso a esta necesidad de que un experimento esté bien diseñado para poder llegar a la conclusión de que hay una relación de causa y efecto. Y el resultado son titulares del estilo de "Los médicos pueden reducir el número de demandas por negligencia médica dedicando más tiempo a sus pacientes". En realidad, lo que se observó es que los facultativos que habían sido menos demandados por negligencia eran los que pasaban mucho tiempo con sus pacientes. Eso no significa que puedas coger a un mal médico, obligarlo a que pase más tiempo con sus pacientes y conseguir que le demanden menos por negligencia; de hecho, dedicarles más tiempo podría incluso crear más problemas.

Capítulo 19

Tablas de contingenciae independencia

En este capítulo

- Elaborar tablas de contingencia con variables categóricas
- Profundizar en las distribuciones marginales, conjuntas y condicionales
- Comprobar la independencia y la dependencia
- Ver los resultados de las tablas de contingencia con perspectiva

Las *variables categóricas* colocan los sujetos en grupos en función de determinadas características, comportamientos o resultados, por ejemplo, si has desayunado esta mañana (sí, no) o cuál es tu filiación política (demócrata, republicano, independiente, "otros"). La gente continuamente busca relaciones entre dos variables categóricas; apenas pasa un día sin que oigas hablar de otra supuesta relación que alguien ha descubierto.

Aquí tienes unos pocos ejemplos que he encontrado en Internet recientemente:

> ✔ Los dueños de perros tienen una probabilidad mayor de llevar su mascota al veterinario que los dueños de gatos.
>
> ✔ El uso frecuente de las redes sociales por parte de los adolescentes está asociado a la depresión.
>
> ✔ Los niños que juegan más con videojuegos sacan mejores notas en las asignaturas de ciencias.

Con toda la información que te dan sobre variables supuestamente relacionadas, ¿cómo sabes qué hay que creer? Por ejemplo, ¿el uso frecuente de las redes sociales causa depresión, o es al revés? O a lo mejor hay una tercera variable relacionada con las otras dos, por ejemplo problemas familiares. En este capítulo te enseño a organizar y analizar datos de dos variables categóricas. Verás cómo se utilizan las proporciones para realizar comparaciones y buscar tendencias generales, y cómo se comprueba la independencia entre dos de ellas. Además, aprenderás a describir adecuadamente relaciones de dependencia, a evaluar resultados que supuestamente indiquen relaciones de causa y efecto, a hacer predicciones y a extrapolar los resultados a una población.

Organizar una tabla de contingencia

Para buscar vínculos entre dos variables categóricas, en primer lugar tienes que organizar los datos que se hayan recopilado, y una estupenda manera de hacerlo es con una tabla. Una *tabla de contingencia* clasifica a las personas en grupos en función de los resultados de dos variables categóricas (por ejemplo, sexo y opinión).

Supongamos que una empresa privada está construyendo un cámping y ha decidido admitir animales de compañía, siempre y cuando vayan con correa. Ahora están intentando decidir si el cámping debería tener una sección aparte para las personas que vayan con animales. Los responsables de la empresa suponen que los campistas que viajen sin mascotas estarán más a favor de habilitar esa sección que los campistas que viajen con mascotas, pero quieren saberlo con certeza. Para ello, seleccionan aleatoriamente 100 campistas de la zona, realizan una encuesta y anotan la opinión de cada uno sobre la conveniencia de habilitar una sección aparte para acampar con mascotas (sí, no), y si ellos mismos acampan con mascota (sí, no). El resultado es una hoja de cálculo con 100 filas de datos, una para cada una de las personas encuestadas. Cada fila tiene dos elementos de datos: una columna donde se indica si el campista viaja con mascota (sí, no) y otra columna donde figura la opinión de esa persona respecto a si debe habilitarse una sección aparte para quienes acampen con mascota (a favor, en contra). Supongamos que las diez primeras filas del conjunto de datos son las que figuran en la tabla 191.

Esta pequeña parte del conjunto de datos ya permite comenzar un análisis. Por ejemplo, a partir de los resultados de la columna 2 sabemos que la mitad de los encuestados (5/10=0,5) van de cámping con mascota, y la otra mitad no llevan mascota. De los campistas que llevan mascota (es decir, de las cinco personas que tienen un "sí" en la columna 2), tres (el 60%) están a favor de que se habilite una sección aparte; y esta proporción es idéntica para los campistas que no llevan mascota. Probablemente los resultados de estos diez campistas no puedan aplicarse a las 100 personas encuestadas; sin embargo, si examinaras manualmente los datos brutos de las 100 filas que integran este conjunto de datos, te costaría muchísimo encontrar tendencias.

Tabla 19-1. Diez primeras filas de datos procedentes de la encuesta a campistas

Persona	*¿Va usted de cámping con mascota?*	*Opinión sobre habilitar una sección aparte para los campistas que lleven mascota*
1	Sí	En contra
2	Sí	En contra
3	Sí	A favor
4	No	A favor
5	No	A favor
6	Sí	A favor
7	No	En contra
8	No	A favor
9	Sí	A favor
10	No	En contra

Para hacerse una idea de lo que está ocurriendo en un conjunto de datos grande al examinar dos variables categóricas, hay que organizar los datos en una tabla de contingencia. En los siguientes apartados te explico cómo hacerlo.

Crear las celdas

Una tabla de contingencia organiza los datos categóricos de dos variables utilizando líneas para representar una variable (por ejemplo, ir de cámping con mascota: sí o no) y columnas para representar la otra variable (por ejemplo, la opinión sobre habilitar una sección aparte para los campistas que viajen con mascota: a favor o en contra). Cada persona aparece exactamente una vez en la tabla.

Continuando con el ejemplo del cámping, en la tabla 192 encontrarás resumidos los resultados de los 100 campistas encuestados.

En la tabla 192 hay 2 × 2 = 4 números. Estos números representan las *celdas* de la tabla de contingencia, cada una de las cuales representa una intersección de una fila y una columna. La celda de la esquina superior izquierda de la tabla representa a los 20 campistas que viajan con mascota y que están a favor de una sección aparte. En la celda superior derecha tenemos los 10 campistas que viajan con mascota y están en contra de una sección aparte. Abajo a la izquierda están los 55 campistas que viajan sin mascota y que están a favor de una sección aparte, y por último, las 15 personas de la celda inferior derecha van de cámping sin mascota y son contrarias a que se habilite una sección aparte.

Tabla 19-2. Tabla de contingencia para los datos de la encuesta a campistas (las 100 filas)

	A favor de una sección aparte para campistas con mascota	*En contra de una sección aparte para campistas con mascota*
Campista con mascota	20	10
Campista sin mascota	55	15

Calcular los totales

Antes de ponernos en serio con el análisis de una tabla de contingencia (en el apartado “Interpretar los resultados de una tabla de contingencia”), calcularemos algunos totales y los añadiremos a la tabla para consultarlos más adelante. Las variables pueden resumirse por separado calculando los *totales marginales*, que representan el número total de cada fila (para la primera variable) y el número total de cada columna (para la segunda variable). Los *totales marginales de fila* forman una columna adicional a la derecha de la tabla, y los *totales marginales de columna* forman una fila adicional en la parte inferior de la tabla.

Por ejemplo, en la tabla 192, que está en el apartado anterior, el total marginal de fila para la fila 1 (el número de campistas que viajan con mascota) es 20 + 10 = 30; el total marginal de fila para los campistas que viajan sin mascota (fila 2) es 55 + 15 = 70. El total marginal de columna para los que están a favor de una sección aparte (columna 1) es 20 + 55 = 75; y el total marginal de columna para los que están en contra de una sección aparte (columna 2) es 10 + 15 = 25.

El *gran total* es la suma de todas las celdas de la tabla, y es igual al tamaño muestral (los totales marginales no se incluyen en el gran total, sino únicamente las celdas). El gran total se pone en la esquina inferior derecha de la tabla de contingencia. En este ejemplo el gran total es 20 + 10 + 55 + 15 = 100. La tabla 193 muestra los totales marginales de fila y columna y el gran total para los datos de la encuesta a campistas.

Los totales marginales de fila siempre suman el gran total, ya que todos los encuestados van de cámping con mascota o sin mascota. En la última columna de la tabla 193 puedes ver que 30 + 70 = 100. De forma similar, la suma de los totales marginales de columna también da siempre el gran total, ya que todos los encuestados están a favor o en contra de que se habilite una sección aparte. En la última fila de la tabla 193 puedes ver que 75 + 25 = 100.

Tabla 19-3. Tabla de contingencia para los datos de la encuesta a campistas, incluidos los totales marginales

	A favor de una sección aparte para campistas con mascota	*En contra de una sección aparte para campistas con mascota*	*Totales marginales de fila*
Campista con mascota	20	10	20+10=30
Campista sin mascota	55	15	30+15=70
Totales marginales de columna	20+55=75	10+15=25	**Gran total=100** (20+10+55+15)

Cuando organices una tabla de contingencia, incluye siempre los totales marginales y el gran total. Eso te ayudará a empezar con buen pie el análisis de los datos.

Interpretar las tablas de contingencia

Después de crear la tabla de contingencia (según lo indicado en el apartado anterior), calculas porcentajes a partir de los datos con el fin de responder a las preguntas formuladas. A continuación planteo algunas preguntas que pueden ser interesantes en relación con el estudio planteado y que puedes responder fijándote en los datos sobre campistas (cada pregunta se aborda en una sección distinta):

- ✔ ¿Qué porcentaje de campistas está a favor de una sección aparte para quienes vayan de cámping con mascota?
- ✔ ¿Qué porcentaje de campistas viaja con mascota y están a favor de habilitar una sección aparte para quienes vayan de cámping con mascota?
- ✔ De todos los campistas que están a favor de habilitar una sección aparte, ¿la mayoría van de cámping sin mascota o con ella?

Estas (y cualesquiera otras) preguntas referidas a los datos se contestan encontrando y trabajando con las proporciones, o porcentajes, de personas que hay en ciertas partes de la tabla. Este proceso requiere calcular y examinar lo que en estadística se llama *distribuciones*. Una distribución, en el caso de una tabla de contingencia, es una lista de todos los resultados posibles para una variable o una combinación de variables, junto con sus proporciones (o porcentajes) correspondientes.

Por ejemplo, la distribución correspondiente a la variable de la mascota muestra los porcentajes de personas que van de cámping con mascota y de las que van sin ella. La distribución correspondiente a la combinación de la variable de la mascota (sí, no) y la variable de opinión (a favor, en contra) muestra los porcentajes de: 1) las personas que van de cámping con mascota y que están a favor de una sección aparte; 2) las personas que van de cámping con mascota y están en contra de una sección aparte; 3) las personas que van de cámping sin mascota y están a favor de una sección aparte; y 4) las personas que van de cámping sin mascota y están en contra de una sección aparte.

En cualquier distribución, todos los porcentajes deben sumar 100%. Si estás utilizando proporciones (decimales), deben sumar 1. Cada persona preguntada debe estar en algún sitio, y no puede estar en más de un sitio a la vez.

En los siguientes apartados te explico cómo encontrar tres tipos de distribuciones. Cada una de ellas ayuda a responder a la pregunta correspondiente de la lista anterior.

Análisis de variables aisladas con distribuciones marginales

Si quieres examinar una única variable en una tabla de contingencia, no debes fijarte en las celdas de la tabla sino en los márgenes. Como hemos visto antes en el apartado "Calcular los totales", los totales marginales representan el número total de cada fila (o columna) por separado. En la tabla de contingencia correspondiente a la encuesta a campistas (mira la tabla 193), los totales marginales para la variable de la mascota (sí/no)

se encuentran en la columna de la derecha, y los totales marginales para la variable de opinión (a favor/en contra) se encuentran en la fila de abajo.

Sin embargo, si quieres hacer comparaciones entre dos grupos (por ejemplo, las personas que van de cámping con mascota frente a las que van de cámping sin mascota), los resultados son más fáciles de interpretar cuando utilizas proporciones en lugar de totales. Supón que se ha encuestado a 350 personas; te resultará más fácil comparar los resultados si te dicen que el 60% está en el grupo A y el 40%, en el grupo B, en lugar de decirte que 210 personas están en el grupo A y 140 están en el grupo B.

Para analizar el comportamiento de una sola de las variables de una tabla de contingencia, hay que encontrar lo que en estadística se llama la *distribución marginal* de esa variable. En los siguientes apartados te explico la forma de calcular una distribución marginal y representarla gráficamente.

Calcular distribuciones marginales

Para encontrar la distribución marginal de una variable de una tabla de contingencia, tienes que dividir el total marginal de cada fila (o columna) por el gran total.

- ✔ Si tu variable está representada por las filas (por ejemplo la variable de la mascota en la tabla 193), pones los totales marginales de fila en los numeradores y el gran total en los denominadores. La tabla 194 muestra la distribución marginal para la variable de la mascota (sí, no).
- ✔ Si tu variable está representada por las columnas (por ejemplo la opinión sobre una sección aparte para mascotas, según se indica en la tabla 193), pones los totales marginales de columna en los numeradores y el gran total en los denominadores. La tabla 195 muestra la distribución marginal para la variable de opinión (a favor, en contra).

En ambos casos, la suma de las proporciones para cualquier distribución marginal debe ser 1 (con redondeo). Todos los resultados de una tabla de contingencia están sujetos a cierto error de redondeo. Para reducir este error, hay que mantener al menos dos dígitos después de la coma decimal.

Tabla 19-4. Distribución marginal para la variable de la mascota

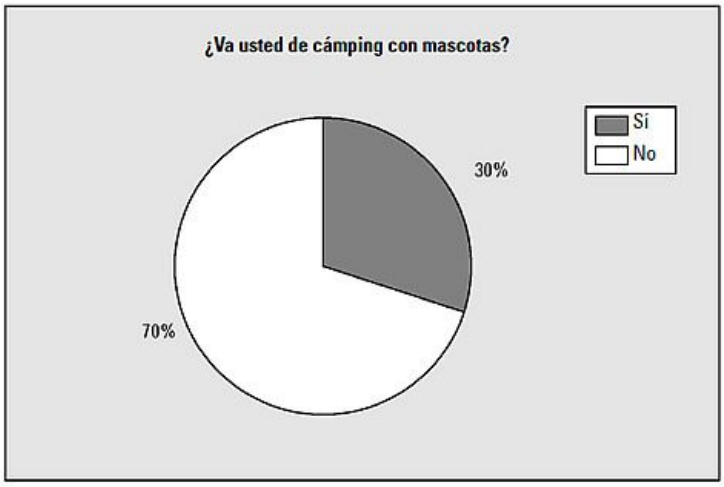
Campista con mascota	*Proporción*
Sí	30 ÷ 100 = 0,3
No	70 ÷ 100 = 0,7
Total	1

Tabla 19-5. Distribución marginal para la variable de opinión

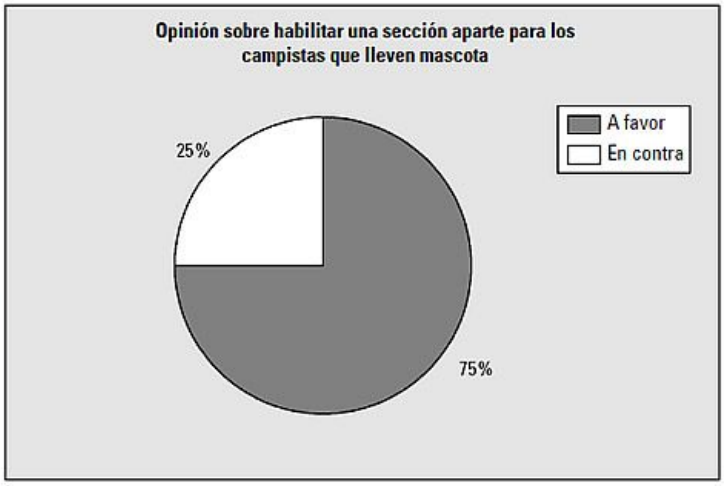
Opinión	*Proporción*
A favor de una sección aparte para campistas con mascota	75 ÷ 100 = 0,75
En contra de una sección aparte para campistas con mascota	25 ÷ 100 = 0,25
Total	1

Representar gráficamente distribuciones marginales

Para representar gráficamente una distribución marginal se puede utilizar un gráfico de sectores o un gráfico de barras. Ambos gráficos muestran la proporción de personas que hay en cada grupo para una única variable. >La figura 191a es un gráfico de sectores que resume la variable de la mascota, y la figura 191b es un gráfico de sectores que desglosa la variable de opinión. Como puedes ver, los resultados de estos dos gráficos de sectores se corresponden con las distribuciones marginales de las tablas 194 y 195, respectivamente.



a



b

Figura 19-1: Gráficos de sectores que muestran las distribuciones marginales para a) la variable de la mascota; y b) la variable de opinión

A partir de los resultados de las dos distribuciones marginales anteriores, una para cada variable, puedes afirmar que la mayoría de las personas de esta muestra van de cámping sin mascota (70%), y que la mayoría de las personas de esta muestra (75%) están a favor de que se habilite una sección aparte para campistas con mascota.

Aunque las distribuciones marginales muestran el comportamiento de cada una de las variables por separado, no nos dicen cuál es la relación existente entre dichas variables. En el ejemplo de los campistas, sabes qué porcentaje de personas está a favor de habilitar una sección aparte, pero no puedes distinguir entre las opiniones de las personas que acampan con mascota y las que acampan sin mascota. Más adelante, en el apartado "Comparar grupos con distribuciones condicionales", explico la manera de realizar este tipo de comparaciones.

Analizar todos los grupos: una distribución conjunta

Permíteme que te cuente una anécdota: cierto fabricante de automóviles realizó una encuesta para averiguar qué características buscan los consumidores en las camionetas *pickup*. Vieron que el color favorito para estas camionetas era el rojo, y la opción más popular era la tracción a las cuatro ruedas. Como respuesta a estos resultados, la empresa empezó a fabricar más camionetas rojas con tracción a las cuatro ruedas.

¿Sabes qué pasó? Metieron la pata hasta el fondo; esas camionetas no se vendían. Resulta que los clientes que compraban camionetas rojas eran en su mayoría mujeres, y las mujeres no elegían la tracción a las cuatro ruedas tanto como los hombres. Los clientes que compraban camionetas con tracción a las cuatro ruedas eran en su mayoría hombres, y en general preferían el color negro antes que el rojo. Por tanto, el resultado preferido de la primera variable (color) combinado con preferido de la segunda variable (opciones del vehículo) no da lugar necesariamente a la combinación más apreciada de las dos variables.

Para saber qué combinación de dos variables categóricas contiene la proporción más alta, tienes que comparar las proporciones de las celdas (por ejemplo, el color y las opciones del vehículo conjuntamente) y no las proporciones marginales (el color y las opciones del vehículo por separado). La *distribución conjunta* de ambas variables en una tabla de contingencia es una relación de todas las combinaciones posibles de filas y columnas junto con la proporción de elementos que hay en cada grupo. Se utiliza para contestar preguntas referidas a dos características; por ejemplo "¿Qué proporción de votantes son demócratas y mujeres?" o "¿Qué porcentaje de campistas viajan con mascota y están a favor de habilitar una sección aparte?". En los siguientes apartados explico la forma de calcular y representar gráficamente una distribución conjunta.

Calcular distribuciones conjuntas

Una distribución conjunta muestra la proporción de datos que hay en cada celda de la tabla de contingencia. En el ejemplo de los campistas, las cuatro combinaciones de filas y columnas son las siguientes:

- ✔ Todos los campistas que viajan con mascota y están a favor de una sección aparte.
- ✔ Todos los campistas que viajan con mascota y están en contra de una sección aparte.
- ✔ Todos los campistas que viajan sin mascota y están a favor de una sección aparte.
- ✔ Todos los campistas que viajan sin mascota y están en contra de una sección aparte.

En todas las proporciones mencionadas en la lista anterior, la frase clave es "todos los campistas". Coges el grupo formado por todos los campistas que han participado en la encuesta y lo divides en cuatro grupos distintos. Cuando veas la palabra "todos", piensa automáticamente que se trata de una distribución conjunta. La tabla 196 muestra la distribución conjunta para todos los campistas que participaron en la encuesta.

Tabla 19-6. Distribución conjunta para los datos de la encuesta a campistas

	A favor de una sección aparte para quienes vayan de cámping con mascota	*En contra de una sección aparte para quienes vayan de cámping con mascota*
Van de cámping con mascota	20 ÷ 100 = 0,2	10 ÷ 100 = 0,1
Van de cámping sin mascota	55 ÷ 100 = 0,55	15 ÷ 100 = 0,15

Para encontrar una distribución conjunta para una tabla de contingencia, coges el número de elementos que hay en cada celda y lo divides por el gran total, para todas las celdas de la tabla. El total de todas estas proporciones debe ser 1 (puede haber cierto error de redondeo).

Para obtener los números que aparecen en las celdas de la tabla 196, coges las celdas de la tabla 193 y divides por el gran total que corresponda (10, en este caso). A partir de los resultados que aparecen en la tabla 196, puedes decir lo siguiente:

- ✔ El 20% de todos los campistas encuestados viajan con mascota y están a favor de una sección aparte (ver la celda superior izquierda de la tabla).
- ✔ El 10% de todos los campistas encuestados viajan con mascota y están en contra de una sección aparte (ver la celda superior derecha de la tabla).
- ✔ El 55% de todos los campistas encuestados viajan sin mascota y están a favor de una sección aparte (ver la celda inferior izquierda de la tabla).
- ✔ El 15% de todos los campistas encuestados viajan sin mascota y están en contra de una sección aparte (ver la celda inferior derecha de la tabla).

Si sumas las proporciones indicadas en la tabla 196, tienes 0,2 + 0,1 + 0,55 + 0,15 = 1. Cada campista aparece en una y solamente una de las celdas de la tabla.

Representar gráficamente distribuciones conjuntas

Para representar gráficamente una distribución conjunta a partir de una tabla de contingencia tienes que elaborar un gráfico de sectores con cuatro sectores, uno para cada proporción de datos correspondiente a una combinación de fila y columna. A los grupos que contienen más elementos les corresponde una porción más grande de la tarta, de manera que tienen más peso cuando se cuentan todos los votos. La figura 192 muestra un gráfico de sectores con la distribución conjunta para los datos de la encuesta a campistas.

En el gráfico de sectores de la figura 192 puedes ver algunos resultados que llaman la atención. La mayoría de las personas de la muestra (0,55 o un 55%) viajan sin mascota y están a favor de que se habilite una sección aparte para campistas con mascota. El sector más pequeño del gráfico representa a las personas que viajan con mascota y están en contra de una sección aparte (0,1 o un 10%).

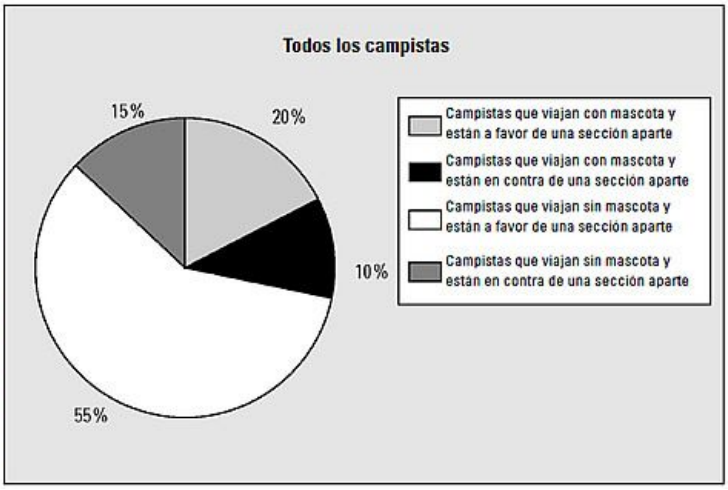


Figura 19-2: Gráfico de sectores que muestra la distribución conjunta de la variable de la mascota y la variable de opinión

Una distribución conjunta proporciona un desglose del grupo completo para las dos variables a la vez, y permite comparar las celdas entre sí y con el grupo entero. Los resultados de la figura 192 revelan que, si hoy mismo se pidiera a todos los encuestados que votaran a favor o en contra de una sección aparte para campistas con mascota, al contar todos los votos se concedería más peso a las opiniones de las personas que van de cámping sin mascota, ya que son la mayoría de las personas encuestadas (el 70%, según la tabla 194), mientras que las personas que van de cámping con mascota tendrían menos voz porque son un grupo más pequeño (30%).

Una limitación de las distribuciones conjuntas es que no permiten comparar de manera equitativa dos grupos entre sí (por ejemplo, las personas que van de cámping con mascota y las que van de cámping sin mascota), ya que este tipo de distribuciones otorgan más peso a los grupos más grandes. En la siguiente sección explico cómo comparar de manera equitativa los grupos de una tabla de contingencia.

Comparar grupos con distribuciones condicionales

Para comparar los resultados de dos grupos (por ejemplo, la opinión de las personas que van de cámping con mascota frente a la opinión de quienes van de cámping sin mascota) necesitas un tipo de distribución diferente de la distribución conjunta. Las *distribuciones condicionales* se utilizan para buscar relaciones entre dos variables categóricas; primero separas las personas en los grupos que quieres comparar (por ejemplo, campistas con mascota y campistas sin mascota), y a continuación comparas los grupos en función de lo que opinan sobre habilitar una sección aparte (sí, no). En los siguientes apartados te cuento cómo se calcula y representa gráficamente una distribución condicional.

Calcular distribuciones condicionales

Para encontrar distribuciones condicionales con fines de comparación, primero tienes que separar las personas en grupos de acuerdo con la variable que quieres comparar. A continuación, para cada grupo, divide el número de personas que hay en una determinada celda por el total marginal correspondiente a ese grupo. Haz esto para todas las celdas de ese grupo. Ahora repite para el otro grupo, utilizando su total marginal como denominador y las celdas de ese grupo como numeradores. (En el apartado anterior "Calcular los totales" encontrarás más información sobre totales marginales.) Ahora tienes dos distribuciones condicionales, una para cada grupo, y puedes comparar de manera equitativa los resultados de los dos grupos.

Para el ejemplo de los datos de la encuesta a campistas (que llevamos observando a lo largo de todo este capítulo), comparas las opiniones de dos grupos: las personas que van de cámping con mascota y las que van sin mascota. En términos estadísticos, quieres encontrar las distribuciones condicionales de opinión tomando como base la variable de la mascota. Así pues, separas los campistas en dos grupos, los que llevan mascota y los que no llevan mascota, y a continuación, para cada grupo, calculas los porcentajes de personas que están a favor y en contra de habilitar una sección aparte. La tabla 197 muestras estas dos distribuciones condicionales en forma de tabla (a partir de los datos de la tabla 193).

Tabla 19-7. Distribuciones condicionales de opinión para campistas con mascota y campistas sin mascota

	A favor de una sección aparte para campistas con mascota	*En contra de una sección aparte para campistas con mascota*	*Total*
Campista con mascota	20 ÷ 30 = 0,67	10 ÷ 30 = 0,33	1
Campista sin mascota	55 ÷ 70 = 0,79	15 ÷ 70 = 0,21	1

La tabla 197 se diferencia de la tabla 196 del apartado anterior "Calcular distribuciones conjuntas" en la manera de contabilizar los valores de la tabla. Ésta es la gran diferencia entre una distribución conjunta y una distribución condicional, y lo que permite realizar comparaciones equitativas utilizando la distribución condicional:

- ✔ En la tabla 196, las proporciones mostradas en las celdas de toda la tabla suman 1 porque el grupo entero se desglosa para las dos variables a la vez en una distribución conjunta.
- ✔ En la tabla 197, las proporciones de cada fila de la tabla suman 1 porque cada grupo se contempla por separado en una distribución condicional.

Representar gráficamente distribuciones condicionales

Una manera efectiva de representar gráficamente distribuciones condicionales es elaborar un gráfico de sectores para cada grupo (por ejemplo, uno para los campistas que viajan con mascota y otro para los que viajan sin mascota), de manera que cada gráfico muestre los resultados de la variable estudiada (opinión: sí o no).

Otro método consiste en utilizar un gráfico de barras apiladas. Un *gráfico de barras apiladas* es un gráfico de barras especial en el que cada barra tiene una altura de 1 y representa un grupo entero (una barra para los campistas que viajan con mascota y otra barra para los campistas que viajan sin mascota). Cada barra muestra el desglose de cada grupo en cuanto a la otra variable estudiada (opinión: sí o no).

La figura 193 es un gráfico de barras apiladas que muestra dos distribuciones condicionales. La primera barra es la distribución condicional de opinión para el grupo de campistas con mascota (fila 1 de la tabla 197) y la segunda barra representa la distribución condicional de opinión para el grupo de campistas sin mascota (fila 2 de la tabla 197).

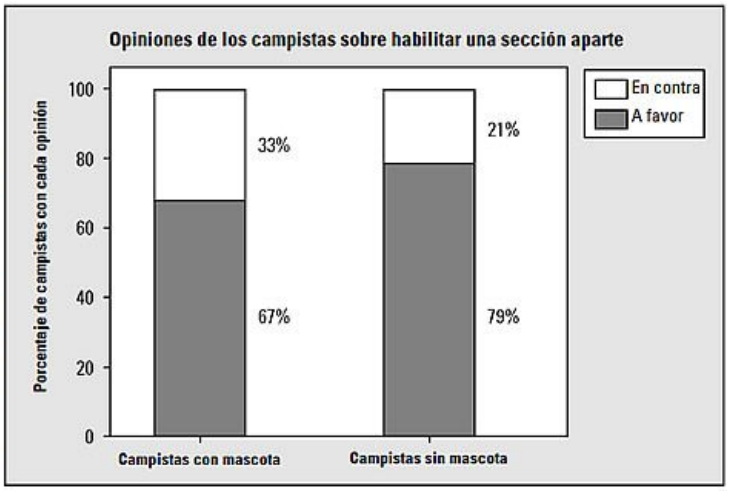


Figura 19-3: Gráfico de barras apiladas que muestra las distribuciones condicionales de opinión para campistas con mascota y campistas sin ella

Usando la tabla 197 y la figura 193, fíjate primero en las opiniones de cada grupo. Más del 50% de los campistas con mascota son partidarios de una sección aparte (el número exacto se redondea al 67%), así que puedes decir que la mayoría de ellos están a favor de una sección aparte. De forma similar, la mayoría de los campistas sin mascota (en torno al 79%, mucho más que la mitad) están a favor de una sección aparte.

Ahora compara las opiniones de los dos grupos cotejando el porcentaje de campistas con mascota que están a favor de una sección aparte (67%) y el porcentaje de campistas sin mascota que están a favor de una sección aparte (79%). Aunque en ambos grupos hay una mayoría de personas que están a favor de habilitar una sección aparte, vemos que el porcentaje de personas partidarias es mayor en el grupo de los campistas sin mascota (porque 79%>67%). Comparando las distribuciones condicionales, has visto que parece existir una relación entre la opinión y el hecho de ir de cámping con mascota. Por tanto, los datos corroboran tu corazonada original, según la cual los campistas de la zona que viajan sin mascota podrían estar más a favor de una sección aparte que los campistas que viajan con mascota.

La diferencia entre resultados encontrada en la figura 193 no es tan grande como habrías podido suponer al ver la distribución conjunta de la figura 192. La distribución condicional tiene en cuenta el número de personas que hay en cada uno de los grupos comparados, mientras que la distribución conjunta pone a todos en el mismo barco. Por eso hay que utilizar distribuciones condicionales para hacer comparaciones equitativas.

A la hora de extraer conclusiones de los datos sobre los campistas, la clave está en comprender que *parece* existir una relación entre las dos

variables contempladas. Los resultados de la encuesta se basan exclusivamente en tu muestra de 100 campistas. Para poder extrapolar estos resultados a toda la población de campistas de la zona que viajan con mascotas o sin ellas (que es lo que realmente quieres hacer), debes tener en cuenta que estos resultados muestrales variarán, y entonces quizá no muestren la misma diferencia. Por eso tienes que realizar un contraste de hipótesis. (Encontrarás toda la información al respecto en el capítulo 14.)

Si quieres realizar un contraste de hipótesis para una relación entre dos variables categóricas (cuando cada variable tiene sólo dos categorías, por ejemplo sí/no o varón/mujer), puedes hacer una prueba para dos proporciones (está en el capítulo 15) o una prueba chi cuadrado (la explico en mi libro *Statistics II for Dummies*, que de momento sólo está disponible en inglés). Si una o más variables tienen más de dos categorías, por ejemplo demócratas/republicanos/otros, tienes que utilizar la prueba chi cuadrado para verificar la independencia en la población.

Puede que te tropieces con un informe en que alguien intente mostrar una relación más fuerte de la que en realidad existe; o al revés: puede que quiera hacer menos obvia una relación mediante la forma de dibujar los gráficos. En los gráficos de sectores, por ejemplo, pueden no indicar el tamaño muestral para hacerte creer que los resultados están basados en una muestra grande, cuando a lo mejor no es así. En el caso de los gráficos de barras, aumentan o reducen la escala para que las diferencias parezcan más grandes o más pequeñas, respectivamente. (En el capítulo 6 encontrarás más información sobre gráficos engañosos para datos categóricos.)

Comprobar la independencia y describir la dependencia

La principal razón de que los investigadores recopilen datos sobre dos variables categóricas es explorar las posibles relaciones entre ellas. Por ejemplo, si una encuesta revela que hubo más mujeres que hombres que votaron por el presidente ganador de las últimas elecciones, concluyes que el sexo y el resultado de los votos están relacionados. Si se encuentra una relación entre dos variables categóricas (es decir, los resultados de los dos grupos son diferentes), los estadísticos dicen que esas variables son *dependientes*.

Sin embargo, si observas que el porcentaje de mujeres que votaron por el presidente actual es idéntico al porcentaje de hombres que votaron por él, entonces las dos variables (sexo y votar por el presidente actual) no están relacionadas y en estadística se dice que son *independientes*. En este apartado explico la forma de comprobar la independencia y describir relaciones dependientes.

Comprobar la independencia

Dos variables categóricas son *independientes* si los porcentajes correspondientes a la segunda variable (que generalmente representa los resultados que quieres comparar, por ejemplo estar a favor o en contra) no difieren de los porcentajes correspondientes a la primera variable (que generalmente representa los grupos que quieres comparar, por ejemplo, hombres frente a mujeres). La independencia puede comprobarse con los métodos explicados en este apartado.

Comparar los resultados de dos distribuciones condicionales

Dos variables categóricas son *independientes* si las distribuciones condicionales son iguales para todos los grupos que se comparan. Las variables son independientes porque, al desglosarlas y compararlas por grupos, los resultados no varían. En el ejemplo de las elecciones que puse al principio del apartado "Comprobar la independencia y describir la dependencia", la independencia significa que la distribución condicional de la opinión es idéntica para los hombres y para las mujeres.

Supongamos que haces la encuesta a 200 votantes para averiguar si el sexo está relacionado con el hecho de que votaran por el presidente actual, y resumes los resultados en la tabla 19-8.

Tabla 19-8. Resultados de la encuesta a votantes

	Votaron por el presidente actual	*No votaron por el presidente actual*	*Totales marginales de fila*
Varones	44	66	110
Mujeres	36	54	90
Totales marginales de columna	80	120	**Gran total = 200**

Para ver si el sexo y el voto son independientes, encuentras la distribución condicional de la orientación de voto de los varones y la distribución condicional de la orientación de voto de las mujeres. Si coinciden, existe independencia; si no, existe dependencia. Estas dos distribuciones condicionales han sido calculadas y aparecen en las filas 1 y 2, respectivamente, de la tabla 19-9. (Encontrarás más información al respecto en el

apartado anterior “Comparar grupos con distribuciones condicionales”.)

Para obtener los números de la tabla 19-9, comencé con la tabla 19-8 y dividí el número de cada celda por su total marginal de fila para obtener una proporción. Cada fila de la tabla 19-9 suma 1 porque cada fila representa su propia distribución condicional (si eres varón, votaste por el presidente actual o no lo hiciste, y lo mismo ocurre con las mujeres).

La fila 1 de la tabla 199 muestra la distribución condicional de la orientación de voto para los varones. Como puedes ver, el 40% de ellos votaron por el presidente actual y el 60%, no. De forma similar, la fila 2 de la tabla muestra la distribución condicional de la orientación de voto para las mujeres; igual que los hombres, el 40% de ellas votaron por el presidente actual y el 60%, no. Como estas distribuciones son idénticas, los hombres y las mujeres votaron igual; el sexo y la orientación del voto son independientes.

Tabla 19-9. Resultados de la encuesta a votantes con distribuciones condicionales

	Votaron por el presidente actual	*No votaron por el presidente actual*	*Total*
Varones	44 ÷ 110 = 0,4	66 ÷ 100 = 0,6	1
Mujeres	36 ÷ 90 = 0,4	54 ÷ 90 = 0,6	1

La figura 194 muestra las distribuciones condicionales de la orientación de voto para varones y mujeres por medio de un gráfico llamado gráfico de barras apiladas. Como las barras son exactamente iguales, concluyes que el sexo y la orientación de voto son independientes.

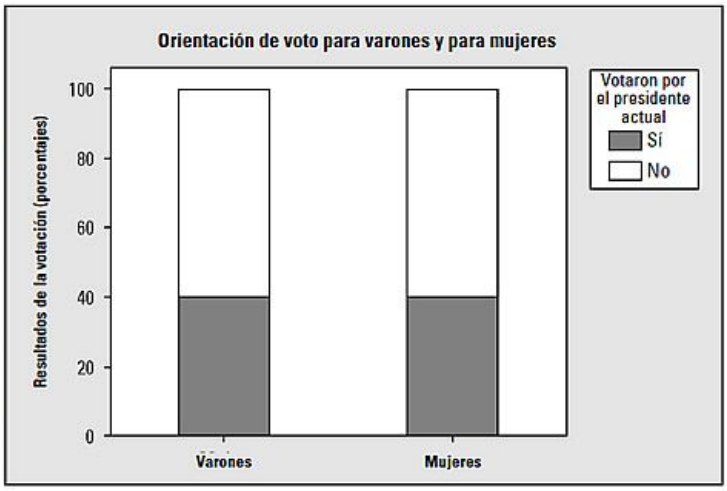


Figura 19-4: Gráfico de barras que muestra las distribuciones condicionales de la orientación de voto para varones y para mujeres

Para que exista independencia, no es necesario que los porcentajes de cada barra sean 5050 (por ejemplo, que el 50% de los varones votaran a favor y el 50% lo hiciera en contra). Lo que debe coincidir no son los porcentajes de cada barra (grupo), sino los porcentajes de barras distintas (grupos distintos) (por ejemplo, que el 60% de los varones esté a favor y el 60% de las mujeres también).

En lugar de comparar las filas de una tabla de contingencia para determinar si existe independencia, también puedes comparar las columnas. En el ejemplo de los votantes, tendrías que comparar el porcentaje de personas de cada sexo que votaron por el presidente actual con el porcentaje de personas de cada sexo que no lo hicieron. La conclusión de independencia sería la misma que obtuvimos anteriormente, aunque los porcentajes calculados serían diferentes.

Comparar distribución marginal y distribución condicional para determinar si existe independencia

Otra manera de comprobar si existe independencia consiste en averiguar si la distribución marginal de la orientación de voto (en general) coincide con la distribución condicional de la orientación de voto para cada uno de los grupos de sexo (varones y mujeres). Si estas distribuciones son iguales, el sexo no importa. De nuevo, el sexo y la orientación de voto son independientes.

En el ejemplo de los votantes, tenemos que la distribución condicional de la orientación de voto para los varones (la primera barra de la figura 194) es el 40% sí y el 60% no. Para encontrar la distribución marginal (global) de la orientación de voto (varones y mujeres juntos), coge los totales marginales de columna que hay en la última fila de la tabla 198 (80 sí y 120 no) y divide por 200 (el gran total). El resultado es 80 / 200 = 0,4 o 40% sí, y 120 / 200 =0,6 o 60% no. (Encontrarás una explicación más detallada en el apartado anterior “Calcular distribuciones marginales”.) La distribución marginal de la orientación de voto en general coincide con la distribución condicional de la orientación de voto para los varones, de manera que la orientación de voto es independiente del sexo.

Ahora es cuando viene de perlas tener una tablita con sólo dos filas y dos columnas. Únicamente tienes que comparar una de las distribuciones condicionales con la distribución marginal, porque únicamente tienes dos grupos que comparar. Si la orientación del voto de los varones coincide con la orientación de voto global, lo mismo ocurrirá para las mujeres. Para comprobar si existe independencia cuando tienes más de dos grupos, hay que utilizar la prueba chi cuadrado. (La explico en mi libro *Statistics II for Dummies*.)

Describir una relación de dependencia

Dos variables categóricas son *dependientes* si las distribuciones condicionales son diferentes para al menos dos de los grupos que se están comparando. En el ejemplo de los votantes presentado en el apartado anterior, los grupos son varones y mujeres, y la variable comparada es si la persona votó por el presidente actual.

La dependencia en este caso consiste en saber que el resultado de la primera variable sí influye en el resultado de la segunda variable. Si se hubiera encontrado dependencia, significaría que los varones y las mujeres no votaron de igual forma (por ejemplo, votaron por el presidente actual más varones que mujeres). Los encuestadores utilizan este tipo de datos como ayuda para controlar mejor sus estrategias electorales.

Otra manera de expresar que dos variables son dependientes es decir que están relacionadas. Sin embargo, los estadísticos no utilizan el término *correlación* para indicar relaciones entre variables categóricas. La palabra *correlación* en este contexto se refiere a la relación lineal entre dos variables numéricas (por ejemplo, altura y peso), como vimos en el capítulo 18 (los medios de comunicación cometen este error continuamente, y los estadísticos nos ponemos de los nervios).

Te pongo un ejemplo para que entiendas mejor la dependencia: el otro día me llamó la atención un comunicado de prensa publicado por el Centro Médico de la Universidad del estado de Ohio. El título decía que la aspirina puede prevenir la aparición de pólipos en personas aquejadas de cáncer colorrectal. Como uno de mis parientes cercanos falleció por culpa de esta enfermedad, me alegré mucho al ver que los investigadores estaban avanzando en ese campo, así que decidí leer el comunicado.

Habían realizado un estudio con 635 pacientes diagnosticados de cáncer colorrectal; aproximadamente la mitad de ellos (317 personas) recibieron un tratamiento con aspirina y la otra mitad (318 personas) tomaron un placebo. Se realizó un seguimiento para ver qué pacientes desarrollaban nuevos pólipos y cuáles no. Los datos del estudio se resumen en la tabla 19-10.

Tabla 19-10. Resumen de los resultados del estudio sobre la aspirina y los pólipos

	Desarrollaron nuevos pólipos	*No desarrollaron nuevos pólipos*	*Total*
Aspirina	54 (17%)	263 (83%)	317 (100%)
Placebo	86 (27%)	232 (73%)	318 (100%)
Total	140	495	635

Comparar los resultados de las filas de la tabla 1910 para comprobar si existe independencia significa contraponer la distribución condicional de los resultados (desarrollar o no nuevos pólipos) para el grupo que recibió el tratamiento con aspirina y la distribución condicional de los resultados para el grupo que tomó un placebo. Si haces los cálculos, obtienes que el 17% (54/317) de los pacientes del grupo tratado con aspirina desarrollaron pólipos, mientras que el resto, un 83%, no los desarrollaron. El 27% (86/318) de los que tomaron el placebo desarrollaron nuevos pólipos y el resto, el 73%, no.

Como el porcentaje de pacientes que desarrollaron pólipos es mucho menor en el grupo de la aspirina que en el grupo del placebo (el 17% frente al 27%), parece existir una relación de dependencia entre el tratamiento con aspirina y el desarrollo de nuevos pólipos por parte de los pacientes de cáncer colorrectal que participaron en este estudio. Pero la pregunta es: ¿pueden extrapolarse estos resultados a toda la población? Encontrarás la respuesta en el apartado "Extrapolar de la muestra a la población", más adelante en este mismo capítulo.

Interpretar los resultados con precaución

Es fácil dejarse llevar al encontrar una relación entre dos variables; en los medios de comunicación ocurre continuamente. Por ejemplo, en un estudio se afirma que comer huevos no influye en el nivel de colesterol como antes se pensaba; al leer el informe con detenimiento ves que el estudio se llevó a cabo con un total de 20 hombres que tenían una salud excelente, seguían dietas bajas en grasa y hacían ejercicio físico varias veces por semana. Diez hombres con buena salud comieron dos huevos al día y su concentración de colesterol no varió apenas en comparación con otros diez hombres que no comieron dos huevos al día. ¿Estos resultados pueden extrapolarse a toda la población? No hay forma de saberlo, porque los sujetos del estudio no representan a todo el mundo. (En el capítulo 17 encontrarás más información sobre cómo evaluar un experimento.)

En este apartado aprenderás a ver los resultados de una tabla de contingencia con la perspectiva adecuada, para tener claro lo que puedes y no puedes afirmar, y por qué. Éste te permitirá evaluar de forma crítica y tomar decisiones sobre los resultados que te presenten (los cuales no son

siempre correctos).

Comprobar si existe una relación de causa y efecto válida

Los investigadores que estudian dos variables generalmente buscan vínculos que apunten a una relación de causa y efecto. Una *relación de causa y efecto* entre dos variables categóricas significa que, si cambias el valor de una variable y todo lo demás permanece igual, eso provoca un cambio en la segunda variable (por ejemplo, que seguir un tratamiento con aspirina reduzca la probabilidad de desarrollar nuevos pólipos en pacientes aquejados de cáncer colorrectal).

Sin embargo, el simple hecho de que dos variables estén relacionadas (sean dependientes) no significa que presenten una relación de causa y efecto. Por ejemplo, observar que las personas que viven cerca de tendidos eléctricos tienen una probabilidad mayor de ir al hospital en el plazo de un año debido a una enfermedad no significa necesariamente que los tendidos eléctricos causaran las enfermedades.

La manera más efectiva de concluir la existencia de una relación de causa y efecto es llevar a cabo un experimento bien diseñado (si es posible). Encontrarás todos los detalles en el capítulo 17, pero ahora mismo puedes leer aquí los puntos más importantes. Un experimento bien diseñado cumple los tres criterios siguientes:

- ✔ Reduce al mínimo el *sesgo* (favoritismo sistemático de sujetos o resultados).
- ✔ Repite el experimento con suficientes sujetos como para que los resultados sean fiables y puedan ser reproducidos por otro investigador.
- ✔ Controla otras variables que puedan afectar al resultado y que no se incluyeron en el estudio.

En el apartado anterior "Describir una relación de dependencia" has visto un estudio en el cual se administró aspirina para prevenir la formación de pólipos en pacientes enfermos de cáncer. Por la manera en que se recopilaron los datos del estudio, puedes confiar plenamente en las conclusiones a las que llegaron los investigadores. Ese estudio es un experimento bien diseñado, según los criterios establecidos en el capítulo 17. Para evitar problemas, los investigadores hicieron lo siguiente:

- ✔ Escogieron aleatoriamente los pacientes que iban a recibir el tratamiento con aspirina y los que tomarían el placebo.
- ✔ Eligieron tamaños muestrales suficientemente grandes para obtener información precisa.
- ✔ Controlaron otras variables eligiendo como sujetos a pacientes que se encontraban en situaciones similares y tenían antecedentes similares.

Como el experimento estaba bien diseñado, los investigadores llegaron a la conclusión de que existía una relación de causa y efecto para los pacientes del estudio. El siguiente paso será comprobar si pueden extrapolar esos resultados a la población formada por todas las personas aquejadas de cáncer colorrectal. De ser así, el título "La aspirina previene el desarrollo de pólipos en pacientes aquejados de cáncer colorrectal" estará plenamente justificado. En el siguiente apartado verás la manera de realizar esta comprobación.

Que se concluya o no la existencia de una relación causal entre dos variables relacionadas depende de la manera en que se haya llevado a cabo el estudio. Un experimento bien diseñado es la manera más convincente de establecer una relación de causa y efecto. En aquellos casos en que realizar un experimento fuera un atentado contra la ética (por ejemplo, obligar a la gente a fumar para demostrar que el tabaquismo causa cáncer de pulmón), haría falta una montaña de estudios de observación convincentes (en los que se recopilan datos sobre fumadores y no fumadores) para demostrar que una asociación entre dos variables pasa a ser una relación de causa y efecto.

Extrapolar de la muestra a la población

En el experimento de la aspirina y los pólipos comentado en el apartado anterior "Describir una relación de dependencia", comparaba el porcentaje de pacientes que desarrollan nuevos pólipos en el grupo tratado con aspirina (el 17%) y en el grupo tratado con placebo (el 27%). Para esta muestra la diferencia es bastante grande, así que, en principio, tengo confianza en que estos resultados podrían extrapolarse a la población formada por todas las personas que tienen cáncer. Pero ¿qué ocurriría si la diferencia entre ambas cifras fuera menor, por ejemplo el 17% frente al 20%, o incluso el 17% frente al 19%? ¿Qué diferencia debe haber entre las proporciones para que pueda suponerse la existencia de una asociación significativa entre las dos variables?

Los porcentajes comparados por medio de datos procedentes de una muestra reflejan las relaciones existentes dentro de esa muestra. No obstante, sabes que los resultados varían de una muestra a otra. Para extrapolar esas conclusiones a la población formada por todos los pacientes de cáncer colorrectal (o cualquier otra población estudiada), la diferencia entre los porcentajes determinados con la muestra debe ser *estadísticamente significativa*. La significación estadística quiere decir que, incluso sabiendo que los resultados varían y teniendo en cuenta esa variación, es muy improbable que las diferencias se deban al azar. De este modo, la misma conclusión extraída sobre una relación puede aplicarse a la población entera, no sólo a un determinado conjunto de datos.

Así pues, cogí los datos del estudio sobre la aspirina y los pólipos y los analicé mediante un contraste de hipótesis para ver la diferencia entre dos proporciones. (Si necesitas repasarlo ve al capítulo 15.) Las proporciones comparadas fueron la proporción de pacientes que recibieron el tratamiento con aspirina y desarrollaron nuevos pólipos y la proporción de pacientes que no tomaron aspirina y desarrollaron nuevos pólipos. A partir de estos resultados, mi valor *p* es inferior a 0,0024 (el *valor p* indica cuál sería la probabilidad de obtener los resultados de la muestra si en realidad no existiera ninguna diferencia entre las poblaciones; en el capítulo 14 encontrarás más información sobre los valores *p*).

Como este valor *p* es muy pequeño, la diferencia entre las proporciones del grupo tratado con aspirina y el grupo tratado con placebo se considera estadísticamente significativa, por lo que concluyo que existe una relación entre tomar aspirina y desarrollar un número menor de nuevos pólipos.

Para extraer conclusiones sobre las relaciones entre variables de una población no puedes basarte exclusivamente en los resultados muestrales de una tabla de contingencia. Debes tener en cuenta el hecho de que los resultados varían de una muestra a otra. El contraste de hipótesis proporciona unos límites para la diferencia que puede existir entre los resultados de la muestra, de forma que todavía pueda afirmarse que las variables son independientes. No te fíes de las conclusiones que se basan exclusivamente en datos muestrales extraídos de una tabla de contingencia.

Realizar predicciones prudentes

Un objetivo común de todas las investigaciones (en particular de los estudios médicos) es realizar predicciones, ofrecer recomendaciones y tomar decisiones después de encontrar una relación entre dos variables categóricas. Sin embargo, como consumidor de información debes ser muy cauto a la hora de interpretar resultados, ya que algunos estudios están mejor diseñados que otros.

El estudio sobre el cáncer colorrectal mencionado en el apartado anterior revela que los pacientes tratados con aspirina presentaron una probabilidad menor de desarrollar nuevos pólipos (el 17%, frente al 27% de los pacientes del grupo placebo). Como este estudio estaba bien diseñado y el contraste de hipótesis para extrapolar los resultados a la población entera fue significativo, es adecuado realizar predicciones y ofrecer recomendaciones para la población de pacientes colorrectales a partir de los resultados muestrales. El título del comunicado de prensa ("La aspirina previene el desarrollo de pólipos en pacientes aquejados de cáncer colorrectal") está totalmente justificado.

Resistir el impulso de extraer conclusiones precipitadas

Cuando alguien afirme que existe una relación entre dos variables categóricas, no te precipites. Tómate tu tiempo para averiguar qué está ocurriendo realmente, incluso si los medios te lo están vendiendo como algo sensacional.

Por ejemplo, mientras escribo estas líneas, en las noticias de una cadena importante de televisión están diciendo que los hombres tienen el 40% más de probabilidades de morir de cáncer que las mujeres. Si eres un hombre, a lo mejor crees que debería entrarte el pánico. Sin embargo, al examinar los detalles, vemos que la realidad es otra muy diferente. Los investigadores observaron que los hombres tardan mucho más que las mujeres en acudir al médico, de manera que cuando les diagnostican un cáncer, está más avanzado y es más difícil de tratar. Como consecuencia, los hombres tienen una probabilidad mayor de morir de cáncer después del diagnóstico (no puede afirmarse que tengan una probabilidad mayor de contraer un cáncer; eso sería el tema de otro estudio). El propósito de este estudio era fomentar la detección precoz como la mejor medida de protección y animar a los hombres a que se hagan chequeos anuales. El mensaje habría quedado más claro si los medios lo hubieran comunicado correctamente (aunque no habría resultado tan emocionante ni tan dramático).

Parte VI

Los decálogos

En esta parte...

Un libro de estadística que se precie no puede dejar de incluir unas cuantas estadísticas propias. En esta parte encontrarás diez formas de convertirte en un sabueso de las estadísticas y diez consejos para sacar mejor nota en un examen de estadística. Esta referencia rápida y concisa te será útil a la hora de criticar o diseñar una encuesta, detectar errores estadísticos comunes y bordar el examen final de tu curso de introducción a la estadística.

Capítulo 20

Diez consejos para ser un sabueso de las estadísticas

En este capítulo

- Detectar los errores estadísticos que suelen cometer los investigadores y los medios de comunicación
- Evitar errores al trabajar con estadísticas

Este libro no sólo pretende ayudarte a comprender las estadísticas que ves a diario en los medios de comunicación y en el trabajo; su principal misión es enseñarte a averiguar si esas estadísticas son correctas, razonables e imparciales. Ante la explosión informativa de hoy en día, debes estar siempre alerta y tener una actitud algo escéptica, porque muchas de las estadísticas que ves están equivocadas o son engañosas, ya sea sin querer o queriendo. Si no valoras de forma crítica la información que consumes, ¿quién crees que lo hará por ti? En este capítulo encontrarás diez consejos para detectar errores estadísticos frecuentes que cometen los investigadores y los medios de comunicación, así como varias maneras de que tú mismo no los cometas.

Detectar gráficos engañosos

La mayoría de los gráficos y diagramas contienen información válida con la que se transmite algo de forma clara, concisa e imparcial. Sin embargo, muchos gráficos proporcionan información incorrecta, mal etiquetada o engañosa, o simplemente no incluyen datos importantes que el lector necesita para tomar decisiones sobre lo que le están presentando. Algunas de estas deficiencias son el resultado de un error, mientras que otras se incorporan de manera intencionada con la esperanza de que los destinatarios no las detecten. Si sabes ver los problemas de un gráfico antes de extraer ninguna conclusión, a ti no conseguirán engañarte. La figura 20-1 muestra ejemplos de cuatro formas de representación gráfica importantes: gráficos de sectores, gráficos de barras, cronogramas e histogramas. En este apartado comento varias maneras de inducirte al error si estos gráficos no se han elaborado correctamente (en los capítulos 6 y 7 encontrarás más información sobre cómo elaborar gráficos y diagramas, y cómo detectar gráficos engañosos).

Gráficos de sectores

Los gráficos de sectores son esquemas circulares (en forma de tarta) divididos en sectores que representan el porcentaje (frecuencia relativa) de elementos pertenecientes a distintos grupos. Los grupos representan una variable categórica, por ejemplo sexo, filiación política o situación laboral. La figura 20-1a muestra un gráfico de sectores con un desglose de las opiniones de los votantes sobre una determinada cuestión (la llamamos "cuestión 1").

Para comprobar si un gráfico de sectores está bien hecho, haz lo siguiente:

- ✔ Asegúrate de que la suma de los porcentajes es el 100% o casi el 100% (el error de redondeo, si lo hay, debe ser pequeño).
- ✔ Mucho ojo si ves un sector del gráfico llamado "Otros"; se utiliza como cajón de sastre. Si este sector es demasiado grande (más grande que otros sectores), el gráfico es demasiado vago. En el otro extremo, los gráficos que tienen un gran número de sectores muy pequeños proporcionan demasiada información.
- ✔ Ten cuidado con las distorsiones propias de los gráficos tridimensionales. El sector que está en primer plano parece más grande de lo que es en realidad debido al ángulo en que se presenta.
- ✔ Busca el número total de elementos que forman el gráfico de sectores, para poder determinar cuál era el tamaño de la muestra antes de que la dividieran en porciones. Si el tamaño del conjunto de datos (el número de personas encuestadas) es demasiado pequeño, la información no es fiable.

Gráficos de barras

Un gráfico de barras es parecido a un gráfico de sectores, salvo que en lugar de tener forma circular y estar dividido en porciones, representa cada grupo como una barra, y la altura de cada una de ellas representa el número (frecuencia) o el porcentaje (frecuencia relativa) de elementos de ese grupo. La figura 20-1b es un gráfico de barras que muestra las frecuencias relativas de las opiniones de los votantes sobre cierta cuestión (la llamamos "cuestión 1"); sus resultados coinciden con los del gráfico de sectores representado en la figura 20-1a.

Cuando examines un gráfico de barras:

- ✔ Comprueba el tamaño muestral. Si las barras representan frecuencias, el tamaño muestral se obtiene sumándolas; si las barras representan frecuencias relativas, tienes que conocer el tamaño muestral para saber cuántos datos se utilizaron para elaborar el gráfico.
- ✔ Ten en cuenta cuáles son las unidades representadas por la altura de las barras y qué significan los resultados. Por ejemplo, ¿muestran

el número total de delitos o la tasa de delincuencia (el número de delitos por habitante)?

- ✔ Mira cuál es el punto inicial del eje donde se encuentran las cantidades (o porcentajes) y fíjate también en los extremos: si las alturas de las barras fluctúan entre 200 y 300 pero el eje comienza en 0, las diferencias parecerán poco importantes. Sin embargo, si el punto inicial del eje es 20, todas las barras estarán cortadas por abajo y entonces las diferencias que haya (entre 0 y 100) parecerán más significativas de lo que deberían.
- ✔ Comprueba el rango de valores en el eje donde se indican las cantidades (o porcentajes). Si las alturas de las barras están entre 6 y 108 pero el eje muestra del 0 al 50, el gráfico tendrá mucho espacio en blanco y las diferencias entre las barras serán difíciles de apreciar. Por el contrario, si el eje va desde 5 hasta 110 y apenas hay espacio en blanco, las barras estarán muy estiradas y las diferencias entre grupos parecerán mayores de lo que son en realidad.

Cronogramas

Un cronograma muestra la variación experimentada por una variable numérica a lo largo del tiempo (por ejemplo la cotización de un valor bursátil, las cifras de venta de un vehículo o la temperatura media). La figura 20-1c es un ejemplo de cronograma que muestra el porcentaje de votos afirmativos entre 2002 y 2010, en incrementos de 2 años.

Cosas que debes tener en cuenta al contemplar un cronograma:

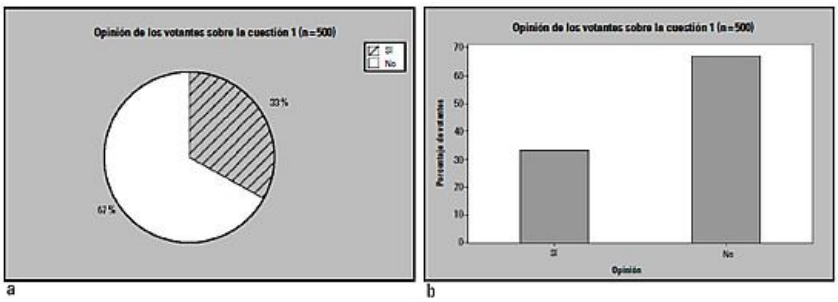
- ✔ Fíjate en la escala del eje vertical (cantidad) y del eje horizontal (línea temporal); simplemente alterando la escala se puede conseguir que los resultados parezcan más o menos significativos de lo que son en realidad.
- ✔ Fíjate en las unidades representadas por el gráfico y asegúrate de que son las adecuadas para realizar comparaciones a lo largo del tiempo; por ejemplo, ¿las cantidades en dólares se han ajustado para tener en cuenta la inflación?
- ✔ No te fíes de las personas que pretendan explicarte la razón de una tendencia sin aportar estadísticas adicionales que respalden su opinión. Un cronograma generalmente muestra algo que está ocurriendo; la causa es algo completamente distinto.
- ✔ Comprueba las distancias de separación entre los puntos del eje temporal. Cuando las separaciones no son uniformes, generalmente es porque faltan datos. Por ejemplo, puedes encontrarte un eje temporal con espacios de igual tamaño entre 2001, 2002, 2005, 2006 y 2008, cuando en realidad deberían haber puesto espacios en blanco en los años para los cuales no hay datos disponibles.

Histogramas

Un *histograma* es un gráfico que divide la muestra en grupos en función de una variable numérica (por ejemplo, edad, altura, peso o ingresos) y muestra el número de elementos (frecuencia) o el porcentaje de elementos (frecuencia relativa) que hay en cada grupo. La figura 20-1d es un histograma de frecuencia que muestra las edades de los votantes en unas determinadas elecciones.

Cosas que debes tener en cuenta al contemplar un histograma:

- ✔ Fíjate en la escala utilizada para el eje vertical (frecuencia/frecuencia relativa). En particular, comprueba si los resultados se han exagerado o disimulado mediante el uso de una escala inadecuada.
- ✔ Comprueba si las unidades del eje vertical indican frecuencias o frecuencias relativas. Si se trata de frecuencias relativas, necesitas el tamaño muestral para saber cuántos datos estás viendo.
- ✔ Fíjate en la escala utilizada para los grupos de la variable numérica en el eje horizontal. Si los grupos están formados a partir de intervalos pequeños (por ejemplo, 0-2, 2-4, etc.), las alturas de las barras posiblemente sean muy variables y dificulten la interpretación del gráfico. Por el contrario, si los grupos se basan en intervalos grandes (por ejemplo, 0-10, 100-20, etc.), puede que la pauta revelada por los datos se vea más uniforme que la pauta real.



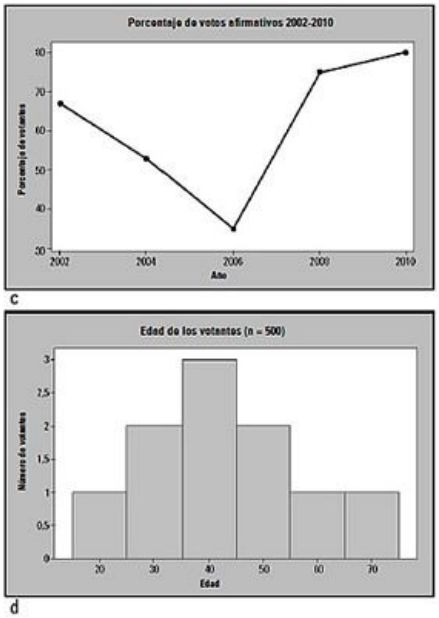


Figura 20-1: Cuatro tipos de gráficos: a) gráfico de sectores; b) gráfico de barras; c) cronograma; d) histograma

Descubrir datos sesgados

El sesgo en la estadística es el resultado de un error sistemático que sobrestima o subestima el valor real. Por ejemplo, si para medir las plantas utilizo una regla que es un centímetro más corta de lo que debería, todos los resultados que obtenga estarán sesgados (serán sistemáticamente menores que los valores reales).

Aquí tienes algunas de las causas más comunes de datos sesgados:

- ✔ Los instrumentos de medición pueden tener un error sistemático. Por ejemplo, el radar portátil de un policía puede decir que ibas a 105 kilómetros por hora cuando tú sabes que no pasabas de 10, o una balanza mal calibrada puede añadir siempre 3 kilos a tu peso real.
- ✔ La manera en que se diseña el estudio puede causar sesgo. Por ejemplo, si en una encuesta se pregunta "¿Alguna vez ha estado usted en desacuerdo con el gobierno?", con toda seguridad se sobrestimará el porcentaje de personas que en general están descontentas con la actuación del gobierno. (En el capítulo 16 explico cómo reducir al mínimo el sesgo en las encuestas.)
- ✔ La muestra de personas puede no ser representativa de la población de interés. Éste sería el caso, por ejemplo, si quieres estudiar los hábitos de estudio de los universitarios y tan sólo vas a la biblioteca de la universidad. (Encontrarás más información en el apartado "Identificar muestras no aleatorias", en este mismo capítulo.)
- ✔ Los investigadores no siempre son objetivos. Imagina que, en el marco de un estudio farmacológico, un grupo de pacientes toma una pastilla de azúcar y el otro grupo toma el fármaco que se quiere analizar. Si los investigadores saben cuál es el grupo experimental, puede que sin querer presten más atención a esos pacientes para ver si el fármaco está surtiendo efecto, y a lo mejor incluso proyectan resultados a los pacientes (por ejemplo diciéndoles: "Seguro que se encuentra usted mejor, ¿verdad?"). Esto crea un sesgo en favor del fármaco. (En el capítulo 17 expongo los criterios que deben seguirse para diseñar un estudio válido.)

Para detectar datos sesgados, averigua cómo se han recopilado los datos. Pregunta cómo se seleccionaron los participantes, cómo se llevó a cabo el estudio, qué preguntas se formularon, qué tratamientos se administraron (medicamentos, técnicas, etc., en su caso) y quién los conocía, qué instrumentos de medición se utilizaron y cómo fueron calibrados, etc. Busca errores de favoritismo sistemáticos, y si los encuentras, no hagas caso de los resultados.

Buscar un margen de error

La palabra "error" tiene una connotación negativa, como si los errores siempre pudieran evitarse. Sin embargo, en estadística no siempre es así. Por ejemplo, cuando alguien intenta estimar un valor poblacional a partir de una muestra distinta de la población entera, es seguro que cometerá, en mayor o menor medida, lo que en estadística se llama un *error de muestreo*. El simple hecho de seleccionar una muestra de la población significa que dejas fuera a cierto número de personas, de manera que es imposible obtener el valor poblacional exacto. En cualquier caso, no hay de qué preocuparse. Recuerda que en estadística no se ofrecen certezas, sino buenas aproximaciones. Si la muestra es lo suficientemente grande, el error de muestreo será pequeño (suponiendo que los datos sean válidos, por supuesto).

Para evaluar un resultado estadístico necesitas una medida de su exactitud, y esta medida suele ser el margen de error. El margen de error te dice cuánto supone el investigador que sus resultados variarán de una muestra a otra. (En el capítulo 12 encontrarás más información sobre el margen de error.) Cuando un investigador o un medio de comunicación no menciona el margen de error, te quedas sin saber si los resultados son exactos, o peor aún, das por sentado que todo está bien, aunque en muchos casos no es así.

Cuando examines resultados estadísticos que impliquen la estimación de un número (por ejemplo, el porcentaje de ciudadanos que están de acuerdo con la actuación de su presidente), mira siempre cuál es el margen de error. Si no te lo dicen, ¡pídelo! (o bien, si dispones de información suficiente, puedes calcularlo tú mismo mediante las fórmulas que aparecen en el capítulo 13).

Identificar muestras no aleatorias

Si quieres estudiar una población entera pero tan sólo puedes estudiar una muestra de individuos, ¿cómo puedes cerciorarte de que esa muestra representa a toda la población? Lo más importante es seleccionar la muestra totalmente al azar, es decir, tomar una *muestra aleatoria*. Una muestra es aleatoria cuando tiene la misma probabilidad de ser elegida que cualquier otra posible de igual tamaño (es como sacar nombres de un sombrero).

Sin embargo, muchas encuestas están basadas en muestras no aleatorias. Por ejemplo, las encuestas en las que se les pide a los televidentes que llamen por teléfono para dar su opinión no representan muestras aleatorias. De hecho, ni siquiera pueden considerarse muestras: cuando tomas una muestra seleccionas a personas de la población, mientras que en las encuestas de llamada voluntaria las personas se seleccionan a ellas mismas. En general, por motivos éticos, los experimentos (en particular los estudios médicos) tampoco pueden llevarse a cabo con una muestra aleatoria de personas. No puedes llamar a alguien y decirle: "Le hemos elegido al azar para que participe en un estudio sobre el sueño. Tendrá que venir mañana a nuestro laboratorio y dormir allí dos noches". Este tipo de experimentos se llevan a cabo con sujetos que se han presentado voluntarios para participar; no han sido seleccionados previamente de forma aleatoria.

No obstante, aunque no puedas seleccionar los sujetos (participantes) del experimento de forma aleatoria, igualmente puedes obtener resultados válidos si incorporas la aleatoriedad por otros medios: asignando los sujetos de forma aleatoria al grupo experimental y al grupo de control. Si los grupos se asignan aleatoriamente, es bastante probable que sean similares, salvo por el tratamiento que reciba cada uno. De este modo, si encuentras una diferencia suficientemente grande en los resultados de los grupos, puedes atribuirla al tratamiento administrado, en lugar de a otros factores.

Antes de tomar ninguna decisión sobre los resultados estadísticos de una encuesta, averigua cómo se seleccionó la muestra de personas. Si la muestra no se seleccionó de forma aleatoria, los resultados deben cogerse con pinzas (como vimos en el capítulo 16). Cuando leas los resultados de un experimento, averigua si los sujetos fueron asignados aleatoriamente al grupo experimental y al grupo de control. Si no fue así, no tengas en cuenta los resultados (mira en el capítulo 17 qué hacer).

El tamaño sí que importa

A la hora de evaluar la precisión de una estadística, tan importante es la calidad como la cantidad de información. Cuantos más datos se hayan analizado, más exacta será la estadística obtenida. La cuestión de la calidad se aborda en el apartado "Descubrir datos sesgados", anteriormente en este mismo capítulo. Una vez que se ha confirmado la calidad de la información, es necesario valorar su precisión, y para eso hay que averiguar cuánta información se ha recopilado (es decir, hay que conocer el tamaño muestral).

Si el tamaño muestral es pequeño, los resultados son menos precisos (a menos que la población también sea pequeña). Muchos titulares de noticias dejan de ser fiables cuando compruebas que el estudio se basó en una muestra pequeña. Y lo que quizá sea peor: muchos estudios no mencionan siquiera el tamaño muestral, lo cual debería hacerte dudar de los resultados (por ejemplo, recuerdo un anuncio de chicles en el que decían: "Cuatro de cada cinco dentistas encuestados recomiendan la marca tal a sus pacientes". ¿Y qué pasa si en realidad sólo preguntaron a cinco dentistas?).

No hace falta que te devanes los sesos con esto, pero para un estadístico (que concede mucha importancia a la precisión) hay una gran diferencia entre decir "4 de cada 5" y decir "4.000 de cada 5.000". Aunque ambas fracciones equivalgan al 80%, la última representa un resultado mucho más preciso (repetible) porque está basada en un tamaño muestral mucho mayor (suponiendo que los datos sean válidos, por supuesto). Si alguna vez te has preguntado cuál es la diferencia entre las matemáticas y la estadística, aquí tienes la respuesta. (En el capítulo 12 se trata más a fondo el tema de la precisión.)

Sin embargo, tener más datos no siempre significa tener mejores datos; eso depende de cómo se recopilaran (ver el capítulo 16). Supongamos que quieres conocer la opinión de los habitantes de una ciudad sobre una propuesta del ayuntamiento. Una muestra aleatoria pequeña con datos bien recopilados (por ejemplo, una encuesta por correo postal a un número reducido de viviendas seleccionadas al azar en un mapa de la ciudad) es mucho mejor que una muestra no aleatoria grande con una mala estrategia de recopilación de datos (por ejemplo, poner una

encuesta en el sitio web del ayuntamiento y pedir a la gente que responda).

Fíjate siempre en el tamaño muestral antes de tomar decisiones sobre información estadística. Cuanto más pequeño sea el tamaño muestral, menos precisa será la información. Si en el artículo no se menciona el tamaño muestral, consigue una copia del informe completo, ponte en contacto con el investigador o ponte en contacto con el periodista que escribió el artículo.

Detectar correlaciones malinterpretadas

Todo el mundo busca conexiones entre variables; por ejemplo, ¿qué grupo de edad tiene una probabilidad mayor de votar al partido demócrata? ¿Si tomo más vitamina C me resfriaré menos? ¿De qué forma afecta a la visión pasar todo el día trabajando delante del ordenador? Cuando piensas en conexiones o asociaciones entre variables, probablemente pienses en correlación. En efecto, la correlación es uno de los estadísticos más utilizados, pero también uno de los más malinterpretados y mal utilizados, sobre todo por parte de los periodistas.

A continuación puedes ver algunos puntos importantes sobre la correlación (en el capítulo 18 encontrarás más información al respecto):

- ✔ **En estadística, la *correlación* (designada con la letra *r*) se define como la medida de la fuerza y la dirección de la relación lineal existente entre dos variables numéricas.** La correlación indica si las variables aumentan a la vez o bien van en direcciones opuestas, y hasta qué punto esta pauta es coherente en todo el conjunto de datos.
- ✔ **El término estadístico "correlación" se utiliza exclusivamente en el contexto de dos variables numéricas (por ejemplo, la altura y el peso).** No se aplica a dos variables categóricas (por ejemplo, sexo y filiación política).

  Por ejemplo, el sexo y la orientación de voto pueden estar relacionados, pero utilizar la palabra "correlación" para describir dicha relación es incorrecto desde el punto de vista de la estadística. Puedes decir que dos variables categóricas están *asociadas*.
- ✔ **Si existe una fuerte correlación entre dos variables numéricas, los puntos del diagrama de dispersión correspondiente deberían revelar una tendencia lineal clara**. Si la distribución de los puntos no se corresponde bien con una recta, probablemente las variables no presenten una correlación *(r)* fuerte, y viceversa. (En el capítulo 18 encontrarás más información sobre rectas de mejor ajuste y modelos de regresión lineal.)

  Una correlación débil implica que no existe una relación lineal entre las dos variables, aunque eso no significa necesariamente que no puedan tener otro tipo de relación. Por ejemplo, las bacterias se multiplican de manera exponencial (su número aumenta cada vez más rápido).
- ✔ **La correlación no significa automáticamente que exista una relación de causa y efecto entre las dos variables.** Por ejemplo, supón que Susana, a partir de sus observaciones, afirma que las personas que beben refrescos*light* tienen más acné que el resto. Aunque seas consumidor habitual de refrescos *light*, no te pongas nervioso todavía. Esta correlación puede ser una simple coincidencia entre las personas que ella observó. Como mucho, significaría que sería necesario hacer más estudios (aparte de la simple observación) para establecer algún tipo de conexión entre los refrescos *light* y el acné. (Susana podría leer el capítulo 17 para ver cómo se diseña un experimento válido.)

Revelar variables de confusión

Una *variable de confusión* es una variable que no se contempla en el estudio pero que puede influir en los resultados y comprometer la validez de las conclusiones. Por ejemplo, imagina que un investigador afirma que comer algas aumenta la longevidad, pero cuando lees el estudio descubres que se basó en una muestra de personas que comen algas habitualmente y tienen más de cien años. Al leer las entrevistas realizadas a esas personas, descubres otros secretos que las han llevado a vivir tantos años (además de comer algas): dormían una media de ocho horas diarias, bebían mucha agua y hacían ejercicio físico todos los días. Entonces, ¿fueron las algas la causa de su longevidad? No hay manera de saberlo, ya que existen variables de confusión (el ejercicio físico, el consumo de agua y los hábitos de sueño) que también pueden haber contribuido.

La mejor manera de controlar las variables de confusión consiste en realizar un experimento bien diseñado (para saber cómo, repasa el capítulo 17), para lo cual hay que seleccionar dos grupos que presenten tantas similitudes como sea posible, salvo que un grupo recibirá un tratamiento específico y el otro grupo será el de control (recibirá un tratamiento ficticio, ningún tratamiento o un tratamiento estándar no experimental). Luego se comparan los resultados de los dos grupos y las diferencias significativas observadas se atribuyen al tratamiento (y a ninguna otra cosa, en un mundo ideal).

El estudio sobre las algas no era un experimento diseñado, sino un estudio de observación. En los estudios de observación no existe ningún tipo de control sobre las variables; simplemente se observa lo que pasa y se anota la información. Los estudios de observación van bien para encuestas y sondeos, pero no para revelar relaciones de causa y efecto, puesto que no tienen en cuenta las variables de confusión. Un experimento bien diseñado proporciona pruebas mucho más sólidas.

En aquellos casos en que llevar a cabo un experimento atenta contra la ética (por ejemplo, demostrar que el tabaquismo causa cáncer de pulmón obligando a la mitad de los sujetos del experimento a fumar diez cajetillas diarias durante veinte años mientras la otra mitad de los sujetos no fuman nada), hay que confiar en las pruebas obtenidas por medio de muchos estudios de observación realizados en muchas situaciones diferentes, todos con el mismo resultado. (En el capítulo 17 encontrarás más información sobre el diseño de experimentos.)

Examinar los números

Sólo porque una estadística aparezca en los medios de comunicación no significa que sea correcta. De hecho, se publican errores continuamente (sin querer o queriendo), así que debes permanecer alerta. Aquí tienes algunos consejos para detectar cifras equivocadas:

- ✔ **Comprueba que los totales cuadran.** En el caso de los gráficos de sectores, asegúrate de que los porcentajes suman 100 (se admite un pequeño error de redondeo).
- ✔ **Revisa incluso los cálculos más elementales.** Por ejemplo, un gráfico de sectores muestra que aproximadamente el 83,33% de los ciudadanos estadounidenses están a favor de una determinada cuestión, pero en el artículo acompañante se dice que son "siete de cada ocho" ciudadanos. ¿Se trata de la misma afirmación en ambos casos? No; 7 dividido entre 8 equivale al 87,5%. El 83,33% son cinco de cada seis.
- ✔ **Fíjate en la tasa de respuesta de una encuesta; no te des por satisfecho conociendo tan sólo el número de participantes** (la tasa de respuesta es el número de personas que han respondido dividido por el número total de personas encuestadas, multiplicado por el 100%). Si la tasa de respuesta es muy inferior al 50%, puede que los resultados estén sesgados porque no sabes qué hubieran dicho las personas que no contestaron. (En el capítulo 16 encontrarás toda la información sobre encuestas y tasas de respuesta.)
- ✔ **Pregúntate si el estadístico utilizado era el apropiado.** Por ejemplo, supongamos que el número de delitos ha aumentado, pero la población también. En lugar de comunicar el número total de delitos, los medios deberían comunicar la tasa de delincuencia (el número de delitos por habitante).

Los estadísticos se calculan mediante fórmulas que extraen los números la información que tú les pides. Las fórmulas no saben si las respuestas finales son correctas o no. Naturalmente, son las personas que están detrás de las fórmulas quienes deben saber lo que se hacen. Quienes no saben lo que se hacen, cometerán errores; y también puede haber quien sí sepa lo que se haga y manipule las cifras con la esperanza de que no te des cuenta. Tú, como consumidor de información (y escéptico confirmado), eres quien debe tomar medidas. Para ello, lo mejor es hacer preguntas.

Informar de los informes selectivos

No puedes dar credibilidad a un estudio en el cual un investigador presente un único resultado estadísticamente significativo pero no mencione sus 25 análisis anteriores, ninguno de los cuales resultó ser significativo. Si hubieras tenido noticia del resto de los análisis, te habrías preguntado si ese único resultado estadísticamente significativo lo es de verdad o bien es fruto del azar (como la idea de que un mono que pulsara aleatoriamente las teclas de una máquina de escribir podría llegar a escribir *El Quijote*). Es una pregunta legítima.

La práctica engañosa de analizar datos hasta encontrar algo es lo que en estadística se llama *rastreo de datos, pesca de datos* o *uso selectivo de datos*. Aquí tienes un ejemplo: supongamos que el sr. García, un investigador, quiere averiguar el motivo de que los niños de primero de primaria discutan tanto entre ellos en el colegio (seguro que no tiene hijos, porque entonces ni se le ocurriría investigar sobre este tema). Diseña un estudio para observar una aula de primero de primaria todos los días durante un mes, y toma nota de todo lo que hacen los niños. Vuelve a su despacho, introduce todos los datos, pulsa un botón para pedir al ordenador que realice todos los análisis habidos y por haber, y se recuesta en su asiento esperando los resultados con impaciencia. Después de todo, con tantos datos seguro que encuentra algo.

Tras escudriñar sus resultados durante varios días, al final le toca la lotería. Sale corriendo de su despacho y le dice a su jefe que tiene que emitir un comunicado de prensa porque ha hecho un descubrimiento sensacional: según su estudio, los alumnos de primero de primaria discuten más cuando 1) el día de la semana termina en "s", o 2) cuando el pez de colores que tienen de mascota en el aula pasa por el boquete del barco pirata hundido en el fondo de la pecera. ¡Felicidades, señor García! Sospecho que, después de pasar un mes entero observando a un grupo de niños de primaria, sus facultades de análisis de datos se han visto algo mermadas.

Lo que quería ilustrar es que, si recopilas suficientes datos y los analizas durante suficiente tiempo, es casi seguro que acabarás encontrando algo, pero ese algo quizá no tenga ningún tipo de significación o se deba a una casualidad imposible de repetir por otros investigadores.

¿Cómo puedes protegerte de resultados engañosos debidos la pesca de datos? Busca más información sobre el estudio, empezando por cuántas pruebas se hicieron en total y cuántas de esas pruebas se consideraron no significativas. Dicho de otro modo: intenta averiguar toda la verdad para así poder ver con perspectiva los resultados significativos.

Para no caer en las redes de los investigadores aficionados al uso selectivo de datos, no te quedes con el primer resultado que oigas, sobre todo si da lugar a una gran noticia o te resulta un poco sospechoso. Ponte en contacto con los investigadores y pídeles más información sobre sus datos, o espera a ver si otros investigadores consiguen verificar y reproducir esos resultados.

Distinguir las anécdotas

Las anécdotas influyen muchísimo en la opinión pública y en el comportamiento de la gente, y sin embargo carecen de toda validez. Una *anécdota* es una historia o un resultado basado en la experiencia o la situación de una única persona. Por ejemplo: La camarera a quien le tocó la lotería... dos veces.

- ✔ El gato que aprendió a ir en bicicleta.
- ✔ La mujer que perdió 40 kilos en dos días con la nueva y milagrosa dieta de la patata.
- ✔ La famosa que dice utilizar el tinte de pelo de supermercado que ella misma anuncia (ya, y yo me lo creo).
- ✔ Las anécdotas proporcionan buenos titulares de prensa; cuanto más sensacionalistas, mejor. Sin embargo, las historias sensacionalistas no son la norma, sino la excepción. No ocurren a la mayoría de las personas.

Quizá creas que a ti las anécdotas no te influyen. ¿Y todas las veces que te has dejado influir por la experiencia de una única persona? Tu vecino está encantado con su proveedor de Internet, de manera que tú decides probarlo también. Tu amigo tuvo una mala experiencia con un coche de una determinada marca, de manera que tú no te molestas siquiera en hacer una prueba de conducción. Un amigo de tu padre murió en un accidente de tráfico porque se quedó atrapado por el cinturón de seguridad, de manera que ahora ya no te lo abrochas nunca.

No pasa nada por tomar algunas decisiones a partir de anécdotas, pero las decisiones importantes deben estar basadas en estadísticas y en datos reales procedentes de estudios bien diseñados e investigaciones fiables.

Una anécdota es en realidad un conjunto de datos con un tamaño muestral igual a uno. No tienes información con la que establecer una comparación, ni estadísticas que analizar, ni posibles explicaciones ni información que seguir, sino tan sólo una historia aislada. No permitas que las anécdotas te influyan mucho. Es mejor que te fíes de estudios científicos y de información estadística basada en muestras aleatorias de personas que representen a sus poblaciones objetivo (no una situación aislada). Cuando alguien intente convencerte de algo contándote una anécdota, sólo tienes que decirle: "¡Enséñame los datos!".

Capítulo 21

Diez consejos para sacar mejor nota en los exámenes

En este capítulo

- Organizar bien los conocimientos
- Desarrollar estrategias inteligentes
- Evitar errores tontos

En todos los años que llevo de profesora (no intentes adivinar cuántos, ¡eso es de mala educación!) he dado clase a más de 40.000 alumnos, y cada uno de ellos ha hecho conmigo un mínimo de tres exámenes. Eso hace un total de 120.000 exámenes que he corregido o ayudado a corregir, y puedes creerme si te digo que he visto de todo. He visto respuestas excelentes, respuestas desastrosas y todas las posibilidades intermedias. Un alumno incluso me escribió una nota en un margen pidiéndome que no fuera muy estricta con él porque se le había escapado el perro y no había tenido tiempo de estudiar. Me he encontrado con respuestas que ni siquiera yo he logrado entender. Me he reído, he llorado y me he sentido orgullosa de cómo han respondido mis alumnos en los exámenes.

En este capítulo he puesto una lista de las diez estrategias más utilizadas por los alumnos que suelen sacar buenas notas. Esos alumnos no son necesariamente más inteligentes que los demás (aunque deben conocer la materia, naturalmente), pero sí están mucho mejor preparados. Como resultado, son capaces de enfrentarse a nuevos problemas y situaciones sin venirse abajo, no suelen cometer esos pequeños errores que acaban bajándote la nota final, y es menos probable que un problema los desconcierte de tal modo que no sepan ni por dónde empezar. Es más probable que sepan resolver el problema (o al menos den una respuesta parcial) porque saben etiquetar la información y organizar el trabajo. De esto no hay duda: la clave para bordar un examen de estadística es la preparación.

Tú también puedes aprobar la asignatura de estadística con buena nota (o con mejor nota todavía, si ya se te da bien la materia) aplicando las sencillas estrategias que explico en este capítulo. Recuerda que todo suma, ¡y empieza hoy mismo a mejorar tu nota de examen!

Averiguar qué es lo que no sabes, y ponerle remedio

Cuando se trata de estadística, distinguir entre lo que sabes y lo que no sabes puede resultar bastante complicado. Lees el libro y entiendes todos los ejemplos de los apuntes, pero no consigues resolver los problemas que te mandan de deberes. Sabes responder a todas las preguntas de estadística que te plantea tu compañero de clase, pero no puedes responder a tus propias dudas. Sales del examen pensando que te ha ido bien, pero luego te dicen la nota y te quedas sorprendido.

¿Qué está ocurriendo? Lo que quiero que entiendas es que, para sacar buena nota, debes ser consciente de lo que sabes y de lo que no sabes. Este conocimiento no es fácil de adquirir, pero merece mucho la pena. Lo habitual es que los alumnos descubran lo que no saben de la peor manera posible: metiendo la pata en un examen. Los errores no son malos de por sí, todo el mundo los comete; lo que importa es cuándo los cometes. Si cometes un error antes del examen, cuando todavía estás a tiempo de descubrir qué es lo que estás haciendo mal, no te costará nada. Si cometes ese mismo error en un examen, te costará puntos.

Una estrategia para descubrir lo que sabes y lo que no sabes consiste en leer detenidamente los apuntes de clase y marcar con un asterisco todo lo que no entiendas. También puedes "examinarte" a ti mismo, como explico más adelante en el apartado "Trampa 'Ya me lo sé' n.º 2", y hacer una lista de problemas que te hayan parecido especialmente difíciles. Luego acude a tu profesor con tus apuntes y tu lista y pídele que repase contigo esas dificultades. Como tus preguntas serán lo suficientemente específicas, tu profesor podrá ir directamente al grano, te dará información y ejemplos concretos y se asegurará de que entiendes un concepto antes de pasar al siguiente. Una reunión de este tipo con tu profesor no tiene por qué durar mucho; a veces te responden una pregunta y se produce una especie de efecto dominó, de manera que de golpe se solucionan varias dudas que también tenías apuntadas.

En lo que respecta a comprender todos los conceptos, ejemplos, fórmulas, notaciones y problemas de deberes antes de hacer el examen, no dejes piedra sin remover. Yo siempre les digo a mis alumnos que treinta

minutos conmigo pueden hacer que su nota mejore el 10%, porque se me da de maravilla explicar cosas y responder preguntas (probablemente mejor que a cualquier compañero, cuñado o amigo que hiciera esta asignatura hace cuatro años con otro profesor). Merece la pena que hagas una visita rápida a tu profesor en su despacho, sobre todo si vas con una lista detallada de preguntas. Si, por alguna razón, tu profesor no está disponible, pídele ayuda a un tutor.

Consejos universales para sacar buena nota

Aquí tienes algunos consejos de carácter general que mis alumnos han considerado útiles:

- ✔ Ya sé que habrás oído esto antes, pero es cierto que tienes mucho ganado si vas a clase todos los días y coges apuntes para luego poder revisarlos. De este modo también te aseguras de que no te has perdido la explicación de algún concepto menos destacado que luego pueda salir en alguna pregunta del examen.
- ✔ No te limites a escribir en el cuaderno lo mismo que ponga el profesor en la pizarra, eso es de aficionados. Los profesionales también apuntan cualquier otra cosa que el profesor destaque especialmente en su explicación, aunque no la escriba en la pizarra. Ésa es la diferencia entre un sobresaliente y un notable.
- ✔ Procura llevarlo todo al día y s erorganizado; así no tendrás que correr cuando se acerque la fecha de entrega. El día que invertí 5 dólares en un buen portaminas, una buena goma, una perforadora de papel y una pequeña grapadora fue uno de lo smejores días de mi época de estudiante. Sí, ya sé que ahora todos esos artículos te costarán 10 dólares o más, pero créeme: merece la pena.
- ✔ Conoce a tu profesor y deja que él o ella te conozca a ti. Si te presentas el primer día le causarás muy buena impresión, y para que se acuerde de tu cara puedes plantearle una duda al terminar la clase (si la tienes, claro) o hacerle alguna visita en horario de atención a los alumnos. No te preocupes por si tus preguntas son tontas; lo que importa no es el nivel donde estás ahora, sino tu voluntad de pasar al siguiente nivel y sacar buena nota en la asignatura. Eso es lo que tu profesor quiere ver.

Evitar las trampas “Ya me lo sé”

¿Qué es una trampa “Ya me lo sé”? Es un término que yo utilizo para esa situación en que dices “Bueno, esto ya me lo sé perfectamente, lo tengo más que controlado”, pero luego llega el examen y ¡zas!, resulta que no te lo sabías tan bien como pensabas y no lo tenías controlado en absoluto. Las trampas “Ya me lo sé” son un peligro porque te hacen pensar que no tienes ninguna duda y que vas a sacar un 10 en el examen, cuando la verdad es que todavía hay cosas que no dominas. Aunque existen varios tipos de trampas “Ya me lo sé”, en este apartado describo las dos más frecuentes y te enseño a evitarlas. Las he llamado “trampa ‘Ya me lo sé’ n.° 1” y “trampa ‘Ya me lo sé’ n.° 2” (muy original, ¿verdad?). Las dos trampas son sutiles y en ellas pueden caer incluso los estudiantes más concienzudos, de manera que no te sientas mal si te ves retratado en este apartado. Simplemente piensa en cuántos puntos te ahorrarás cuando salgas del modo “Ya me lo sé” y entres en el modo “Espera un momento, aquí hay algo que no termino de entender bien”.

Trampa “Ya me lo sé” n.° 1

La trampa “Ya me lo sé” n.° 1 ocurre cuando repasas los apuntes de arriba abajo mientras dices: “Vale, esto ya me lo sé”, “Esto lo tengo claro”, “Está chupado”, pero no intentas hacer los problemas desde cero por ti mismo. Si entiendes un problema que ha hecho otra persona, sólo significa que entiendes lo que hizo esa persona cuando resolvió el problema. No quiere decir que tú puedas resolverlo por ti mismo en un examen, cuando estás bajo presión y tienes los ojos clavados en un espacio en blanco donde se supone que va tu respuesta. ¡Hay una gran diferencia!

Yo también caigo a veces en la trampa “Ya me lo sé” n.° 1. Hace poco me leí de cabo a rabo el manual del grabador de vídeo digital y lo tenía todo clarísimo, pero una semana más tarde quise grabar una película y no supe ni por dónde empezar. ¿Por qué no? Había comprendido todo lo que decía el manual pero no había intentado ponerlo en práctica, y cuando llegó el momento no conseguí recordar cómo se hacía.

Los alumnos siempre me dicen que, si ven un problema resuelto, luego ya lo entienden. La pega es que casi todo el mundo es capaz de entender un problema que alguien ya ha resuelto antes. De hecho, la clave está en saberlo razonar y resolverlo sin ayuda, como ocurre en un examen.

Para evitar la trampa “Ya me lo sé” n.° 1, busca en tus apuntes algunos ejemplos que utilizara tu profesor y escríbelos en hojas aparte (sólo el problema, no la solución). A continuación, mezcla todas las hojas y conviértelas en un examen. Para cada problema, empieza escribiendo tan sólo el primer paso. No te preocupes por terminar los problemas; simplemente empiézalos. Una vez hayas hecho esto con todos los problemas, vuelve a mirar los apuntes para ver si los has empezado correctamente (en el reverso de cada hoja, anota de qué parte de los apuntes proviene, para luego tardar menos en comprobar las respuestas).

Trampa “Ya me lo sé” n.° 2

La trampa “Ya me lo sé” n.° 2 es incluso más sutil que la trampa n.° 1. Un alumno entra en mi despacho después del examen y me dice: “Repasé todos los problemas de los apuntes, volví a hacer los problemas de los deberes e incluso hice los exámenes que puso usted de ejemplo, y no cometí apenas errores. Pero a pesar de toda esa preparación, el examen me ha ido de pena”.

¿Qué ha ocurrido? Nueve de cada diez veces, los alumnos que caen en la trampa “Ya me lo sé” n.° 2 hacen de verdad todos esos problemas y se pasan horas y horas estudiando. Lo que pasa es que, cuando se quedan atascados con un problema, echan un vistazo a la solución (que tienen

justo al lado, bien a la vista), averiguan en qué se han equivocado, se dicen a sí mismos: "Ah, claro, qué error más tonto, eso ya lo sabía" y continúan. Al final creen que han hecho todos los problemas bien ellos solos, pero el día del examen pierden algunos puntos, o muchos, en función de dónde se quedaran atascados cuando estaban estudiando.

Entonces, ¿cómo se puede evitar la trampa "Ya me lo sé" n.° 2? Haciendo un examen de prueba en condiciones "reales", con toda la presión. Te explico cómo:

1. **Estudia tanto como te haga falta, de la manera que te parezca más adecuada, hasta que te sientas preparado para poner a prueba tus conocimientos.**

2. **Siéntate a la mesa con un examen de práctica. Si no te lo proporciona el profesor, hazte uno tú mismo eligiendo unos cuantos problemas de los deberes, de los apuntes o del libro de texto.**

   Igual que en un examen de verdad, necesitarás un lápiz, una calculadora y los materiales que te permitan llevar (¡y nada más!). El hecho de no poder consultar el libro ni los apuntes quizá te ponga un poco nervioso o te haga sentir desprotegido, pero es imprescindible que averigües lo que puedes hacer por ti mismo antes de que llegue la hora de la verdad.

   Algunos profesores dejan llevar al examen una *hoja-resumen* (también llamada *chuleta*) en la que puedes escribir cualquier información que consideres útil, dentro de unos límites. Si te dejan llevar hojas-resumen a los exámenes, utiliza una también en el examen de práctica.

3. **Pon en marcha el cronómetro (o el contador del horno) con el mismo tiempo que te vayan a dar para el examen real y ponte manos a la obra.**

4. **Resuelve todos los problemas que puedas lo mejor que sepas y, cuando termines (o se te acabe el tiempo), deja el lápiz sobre la mesa.**

5. **Cuando hayas terminado el "examen", ponte en la posición del loto, inspira profundamente, aguanta el aire unos instantes y expúlsalo en tres veces. A continuación mira las soluciones y corrige la prueba igual que lo haría tu profesor.**

Si ha habido algún problema que no hayas sabido empezar siquiera (aunque sea por alguna tontería y luego, al ver las soluciones, recuerdes de inmediato lo que habías olvidado), entonces no puedes decir: "Ah, sí, esto lo sabía; en un examen de verdad no cometería este fallo"; tienes que decir: "No he sabido empezar el problema sin ayuda. En un examen de verdad me lo habrían contado mal. Tengo que repasarlo bien para que no me vuelva a ocurrir".

En un examen de verdad no tienes una segunda oportunidad. Por lo tanto, cuando estés estudiando no tengas miedo de admitir que no sabes hacer un problema por ti mismo; simplemente alégrate de haberte dado cuenta a tiempo, y ponle remedio. Repasa los apuntes, mira en el libro de texto, pregunta a tu profesor, haz otros problemas del mismo tipo o pide a un compañero de clase que te pregunte. Además, intenta encontrar una pauta en el tipo de problemas que te hayan resultado más complicados o no hayas sabido resolver. Averigua por qué fallaste donde fallaste. ¿Leíste las preguntas demasiado rápido y por eso te equivocaste en las respuestas? ¿Fue un problema de vocabulario o de notación? ¿Te habías estudiado todas las preguntas del examen?...

Cuesta mucho ser crítico con uno mismo y admitir que no sabías algo que pensabas que sí sabías. Pero si coges el toro por los cuernos y descubres tus errores antes de que te cuesten puntos, podrás centrarte en tus puntos débiles y convertirlos en puntos fuertes, llegarás a dominar de verdad la materia y conseguirás una nota más alta en el examen.

Hacerse amigo de las fórmulas

Muchos alumnos se sienten incómodos con las fórmulas (a menos que estén chiflados por las matemáticas, en cuyo caso las ven como una fuente de diversión constante). Este malestar es comprensible —a mí también solían intimidarme (me refiero a las fórmulas, no a los chiflados por las matemáticas)—. El problema es que en una clase de estadística no se puede sobrevivir mucho tiempo sin utilizar fórmulas, de manera que es importante acostumbrarse a ellas desde el principio. Una fórmula te dice mucho más que la forma de calcular algo. Muestra el razonamiento que subyace a los cálculos. Por ejemplo, puedes obtener una visión general de la desviación estándar analizando su fórmula:

$$s=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$$

Si a un valor del conjunto de datos, x_i, le restas la media, $\bar{x}$, sabrás a qué distancia se encuentra ese número por encima o por debajo de la media. Como no quieres que las diferencias positivas y negativas se cancelen entre sí, elevas todas al cuadrado para hacerlas positivas (pero recuerda

que con esto obtienes unidades al cuadrado). A continuación las sumas todas y divides por n–1, que es parecido a encontrar una media, y sacas la raíz cuadrada para regresar a las unidades originales. En un sentido general, estás encontrando algo parecido a la distancia media respecto de la media.

Si retrocedes más aún, a partir de la fórmula sabes que la desviación estándar no puede ser negativa, porque todo está elevado al cuadrado. También sabes que el valor más pequeño que puede adoptar es cero, cosa que ocurre cuando todos los datos son el mismo (es decir, todos son iguales a la media). Asimismo, puedes ver que los datos que están alejados de la media aumentan el valor de la desviación estándar, y viceversa.

Todo este análisis tiene una utilidad concreta. Como ahora entiendes la fórmula de la desviación estándar, sabes lo que está midiendo en realidad: la dispersión de los datos en torno a la media. Así pues, cuando en una pregunta de examen te pidan la dispersión en torno a la media, sabrás exactamente qué hacer. ¡Toma ya!

Para sentirte cómodo con las fórmulas, sigue los siguientes consejos:

- ✔ **Ten la actitud mental correcta.** Piensa que las fórmulas no son más que una forma de sintetizar los cálculos matemáticos. Sólo tienes que saber descifrarlas. Muchos profesores te dejarán llevar una hojaresumen al examen o directamente te proporcionarán una hoja de fórmulas, así que a lo mejor no tienes siquiera que aprendértelas de memoria.
- ✔ **Entiende todos los detalles de todas las fórmulas.** Para que una fórmula te sea de utilidad, es importante que entiendas todos sus componentes. Por ejemplo, antes de utilizar la fórmula de la desviación estándar debes saber qué son x_i y $\bar{x}$, y cuál es el significado de $\sum_{i=1}^{n}$. De lo contrario, la fórmula no sirve de nada.
- ✔ **Practica con las fórmulas desde el primer día.** Utilízalas para comprobar los cálculos hechos en clase o en el libro de texto. Si te sale una respuesta diferente de la que pone, averigua qué has hecho mal. No pasa nada por equivocarte ahora; has detectado el problema a tiempo y eso es lo que cuenta.
- ✔ **Cuando apliques una fórmula para resolver un problema, primero escríbela y luego sustituye las letras y símbolos por números.** Cuantas más veces escribas una fórmula, más cómodo te sentirás cuando tengas que usarla en un examen. Y si, esperemos que no, escribes mal la fórmula, el profesor podrá ver cuál ha sido tu error y posiblemente te cuente algo por esa pregunta.

Si en clase has aprendido algunas fórmulas, lo más seguro es que luego tengas que utilizarlas en el examen. No esperes poder utilizar esas fórmulas con plena confianza en el examen si no has practicado antes con ellas y las has escrito muchas, muchas veces. Empieza con problemas fáciles, así cuando la cosa se complique no tendrás que preocuparte tanto.

Elaborar una tabla procedimental

En el fútbol americano, los *quarterbacks* siempre dicen que intentan ver el juego como si todo ocurriera a cámara lenta, para así tener más tiempo de pensar y reaccionar. En un examen de estadística ocurre exactamente lo mismo (ya ves, tienes algo en común con las estrellas de la Super Bowl). El *quarterback* empieza a ver las cosas a "cámara lenta" cuando empieza a distinguir pautas en la formación defensiva del equipo contrario, en lugar de ver todas las jugadas diferentes. De forma similar, en un examen tú también entras a cámara lenta cuando clasificas los problemas en categorías a medida que los vas leyendo, en lugar de pensar que no se parecen a nada de lo que hayas visto antes.

Para lograr esto, a muchos de mis alumnos les ayuda elaborar una tabla procedimental. Una *tabla procedimental* sirve para clasificar los tipos de problemas que probablemente te encuentres según las estrategias para resolverlos, con ejemplos que puedas consultar rápidamente. La idea básica es organizar lo que sabes siguiendo el hilo de pensamiento siguiente: "Si el problema me pide *X*, lo resuelvo haciendo *Y*, de esta manera". Una tabla procedimental tiene tres columnas:

- ✔ **Pregunta.** En esta columna tienes que escribir una breve descripción de lo que te piden que hagas. Por ejemplo, si el problema te pide que contrastes una afirmación referida a la media poblacional (en el capítulo 14 encontrarás más información al respecto), escribe "Contrastar una afirmaciónmedia poblacional". Si te pide una estimación de la media poblacional (más información en el capítulo 13), escribe "Estimar media poblacional".

Los problemas pueden enunciarse de distintas formas, porque así es cómo funciona el mundo real. Si te encuentras con enunciados diferentes para un mismo problema, debes ponerlos todos en la misma celda, puesto que piden la misma cosa. Por ejemplo, puede que un problema te pregunte directamente una estimación de la media poblacional, mientras que otro diga: "Indica un rango de

valores probables para la media poblacional". Ambas preguntas piden lo mismo, de manera que tienes que incluirlas en la misma celda.

- ✔ **Estrategia.** En esta columna tienes que poner el procedimiento estadístico, fórmula o técnica concreta que necesitas para resolver cada tipo de problema. Para ello, utiliza la jerga técnica. Por ejemplo, si en la primera columna has escrito "Contrastar una afirmaciónmedia poblacional", en esta segunda columna deberías poner "Contraste de hipótesis para μ". Si en la primera columna has escrito "Estimar una media poblacional", en la segunda columna debes poner "Intervalo de confianza para μ".

A la hora de asignar estrategias a las distintas preguntas, guíate por los ejemplos incluidos en tus apuntes de clase y en el libro de texto.

- ✔ **Procedimiento.** En la tercera columna escribe un ejemplo, una fórmula o una nota breve que te oriente en la dirección correcta. Escribe cualquier cosa con la que te sientas cómodo (no va a verlo nadie más que tú, así que hazlo a tu manera). Por ejemplo, si en la primera columna has escrito "Estimar la media poblacional" y en la segunda columna has escrito "Intervalo de confianzamedia poblacional". En la tercera columna puedes poner la fórmula.

Aunque acabo de darte una explicación muy larga y detallada, en realidad hacer una tabla procedimental es la mar de sencillo. A continuación, a modo de ejemplo, puedes ver cómo se introduciría en una tabla procedimental el problema del intervalo de confianza que acabo de mencionar.

Pregunta	*Estrategia*	*Procedimiento*
Estimar la media poblacional (también se llama rango de valores probables)	IC para μ	$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$

Utilizando estas tres columnas, continúa introduciendo en tu tabla procedimental los distintos tipos de problemas que te hayan explicado en clase. No hace falta que escribas hasta el último ejemplo; busca modelos recurrentes y reduce la lista de problemas a un número razonable.

Personaliza la tabla procedimental según tus necesidades particulares, porque sólo te servirá de algo si la haces tú mismo. Cada persona es un mundo, y lo que le sirve a tu amigo quizá no te sirva a ti. No obstante, a lo mejor te viene bien comparar tu tabla con la de un compañero de clase para ver si te has dejado algo.

Si te dejan hacer el examen con una hojaresumen, te sugiero que en una cara pongas tu tabla procedimental. En la otra cara, anota las pequeñas perlas de información que el profesor os dio en clase pero no escribió en la pizarra. Y aunque no te dejen llevar una hojaresumen, yo te aconsejo que igualmente prepares una y la utilices para estudiar. Te ayudará a tener las ideas claras, y así te costará menos ver qué te piden en los problemas y cómo tienes que resolverlos. Si sales del examen sin haber mirado siquiera la hojaresumen, significa que has hecho bien las cosas: mientras escribías la hoja fuiste interiorizando todos los contenidos.

Entender qué te están preguntando

Mis alumnos me dicen a menudo que no entienden lo que pide el problema. Ésa es la pregunta del millón de dólares, ¿verdad? Y no es un asunto en absoluto trivial. La pregunta suele estar oculta en el enunciado del problema; generalmente no es directa como "Encuentra la media de este conjunto de datos".

Por ejemplo, es posible que te pidan que "interpretes" un resultado estadístico. Pero ¿qué significa "interpretar"? Para la mayoría de los profesores, "interpretar" significa explicar algo con palabras que pueda entender cualquier persona sin conocimientos de estadística.

Supongamos que te dan un informe donde se analizan los datos sobre el número de delitos y el número de policías de una determinada ciudad, y te piden que interpretes la correlación existente entre esas dos variables. En primer lugar, encuentras la cifra que corresponde a la correlación (supongamos que es –0,85) y a continuación hablas de sus características más importantes con un lenguaje que sea fácil de entender. La respuesta que yo querría ver en un examen sería algo parecido a lo siguiente: "La correlación entre el número de policías y el número de delitos es –0,85, lo cual indica un relación lineal negativa muy marcada. Cuando el número de policías aumenta, el número de delitos disminuye".

Si entiendes qué te está pidiendo el problema, es más probable que sepas resolverlo. Cuando sabes lo que tienes que hacer, te sientes más seguro. Por el contrario, si no entiendes lo que pide el problema, incluso comenzarlo te parecerá muy difícil. Eso hará que te pongas más nervioso, lo cual, a su vez, puede influir en tu capacidad de resolver el resto de los problemas. Entonces, ¿qué hay que hacer para interpretar correctamente un enunciado? Aquí tienes algunos consejos:

- ✔ **Lee despacio la última frase del problema. Ahí es donde suele estar la pregunta**. En lugar de leer todo el enunciado una segunda (y tercera, y cuarta) vez y ponerte más nervioso todavía, léelo una sola vez y concéntrate en la parte final.
- ✔ **Practica antes del examen.** Repasa todos los ejemplos de los apuntes, los deberes y los problemas del libro de texto, e intenta dilucidar qué es lo que se pregunta en cada caso. Al final empezarás a distinguir pautas en la redacción de los problemas, y cada vez te costará menos ver lo que te piden en realidad.
- ✔ **Pregúntale al profesor qué pistas tienes que buscar y llévale algunos problemas de ejemplo.** Cuando vea que intentas obtener una visión más amplia, lo dejarás impresionado (eso es algo que a los profesores nos encanta). Luego puedes incorporar esos ejemplos a tu tabla procedimental. (Releer el apartado "Elaborar una tabla procedimental".)
- ✔ **Traduce el enunciado del problema a un enunciado estadístico.** Para eso tienes que etiquetar la información que te dan (como explico en el siguiente apartado) y también la que quieres encontrar.

  Por ejemplo, el profesor Lucio quiere ponerle un sobresaliente al 20% de sus alumnos en el examen de estadística, y tú tienes que averiguar cuál es la nota de corte para sacar un sobresaliente, lo cual es lo mismo que decir: "Encuentra la nota que corresponde al percentil 80".

Etiquetar la información que te dan

Muchos alumnos intentan resolver los problemas lanzándose directamente a hacer cálculos con los números que les dan. Este sistema puede funcionar con los problemas más sencillos, pero llega un momento en que deja de valer, y entonces necesitas un método más estructurado. En este sentido, el hábito de etiquetarlo todo correctamente te será muy útil, ya que la etiqueta es lo que permite establecer la conexión crítica entre la primera y la segunda columna de la tabla procedimental (descrita anteriormente en este mismo capítulo). Puede que leas un problema y sepas lo que tienes que hacer, pero si no sabes cómo utilizar la información que te dan, no podrás resolverlo correctamente. Para entender de verdad los números que te da el problema, escribe junto a cada uno lo que significa.

Supongamos que te ponen el siguiente problema: "Tienes un piso en una determinada ciudad y quieres predecir su precio (en miles de euros) a partir de su superficie útil (en metros cuadrados). Recopilas datos sobre 100 casas seleccionadas al azar que se han vendido recientemente. Compruebas que el precio medio es de 219.100 euros, con una desviación estándar de 60.100 euros, y sabes que la superficie media es de 92 metros cuadrados, con una desviación estándar de 16 metros cuadrados. Calculas que la correlación entre superficie y precio para estas casas es de +0,9. Encuentra la recta de regresión de mejor ajuste para predecir el precio de la casa a partir de su superficie".

El primer paso es etiquetarlo todo. Como te basas en la superficie para predecir el precio, la superficie debe ser la variable *x* y el precio debe ser la variable *y*. Las medias son $\bar{x}$ = 92 (metros cuadrados) e $\bar{y}$ = 219,1 (miles de dólares); las desviaciones estándares son s_x = 16,61 (metros cuadrados) y s_y = 2,85 (euros), y la correlación es r = 0,90. El tamaño muestral es n = 100. Ahora ya puedes poner los números en las fórmulas que corresponda. (En el capítulo 18 tienes la información sobre la correlación y la regresión.)

Si sabes que vas a trabajar con una recta de regresión y que vas a utilizar fórmulas, resulta muy útil tener toda la información perfectamente organizada y etiquetada. Así tienes una cosa menos en que pensar. (Este problema en concreto está resuelto en el apartado "Establecer la conexión y resolver el problema".) Si este ejemplo no te convence, aquí tienes otras seis razones para etiquetar toda la información que te den en un problema:

- ✔ **Las etiquetas facilitan el repaso.** Cuando repases el problema (como te aconsejo en el apartado "Haz las cuentas... dos veces"), te resultará muy sencillo ver el hilo del razonamiento que seguiste para resolverlo.
- ✔ **Tu profesor quedará impresionado.** Verá las etiquetas y sabrá que al menos entiendes lo que significa la información que te han dado. De este modo, aunque te equivoques en los cálculos puede que te den algún punto por esa pregunta.
- ✔ **Las etiquetas ahorran tiempo.** Ya sé que escribir más información parece una extraña manera de ahorrar tiempo, pero las etiquetas permiten identificar la información necesaria en un santiamén.

  Por ejemplo, supongamos que tienes que calcular un intervalo de confianza del 95% para la media poblacional (del modo descrito en el capítulo 13) y te dicen que la media muestral es 60, la desviación estándar de la población es 10, y el tamaño muestral es 200. Sabes que en la fórmula deben aparecer $\bar{x}$, σ y n, y conoces una que los incluye a todos:

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

Como ya lo tienes todo etiquetado, simplemente coges lo que necesitas, lo metes en la fórmula, utilizas un valor z^* de 1,96 (el valor crítico correspondiente a un nivel de confianza del 95%) y haces los cálculos para obtener la respuesta:

$$112{,}85 \pm 1{,}96 \frac{10}{\sqrt{200}} = 112{,}85 \pm 1{,}39$$

- ✔ **Las etiquetas ayudan a organizar la mente.** Si trabajas con símbolos, y no sólo con números, es menos probable que te pierdas en los cálculos y olvides lo que estás haciendo. Si clasificas toda la información que te dan, no tendrás que leer el problema una y otra vez, y no te pondrás tan nervioso.
- ✔ **Las etiquetas ayudan a decidir qué fórmula o técnica vas a utilizar para resolver el problema.** Por ejemplo, si crees que debes hacer un contraste de hipótesis pero no te dan una afirmación sobre la media poblacional, significa que no vas bien encaminado. Posiblemente tengas que calcular un intervalo de confianza. Darte cuenta de esto te ahorra un tiempo precioso, ya que no avanzas en la dirección equivocada. Las etiquetas ayudan a reducir rápidamente el abanico de opciones.
- ✔ **Las etiquetas ayudan a resistir el impulso de llenarlo todo de números sin más.** Si te lanzas a hacer cálculos, es más probable que acabes dando una respuesta incorrecta y no te cuenten ningún punto por el problema. Quizá tu profesor no pueda seguir tu razonamiento, o sencillamente no esté dispuesto a invertir todo el tiempo que necesitaría para ello (lo siento, pero esto ocurre a veces).

Poner etiquetas te ahorrará nervios, tiempo y puntos cuando hagas un examen. Sin embargo, la clave está en practicar desde el principio, cuando los problemas todavía son fáciles. Si nunca te has puesto a clasificar la información de un problema, no esperes que el día del examen sepas hacerlo por ciencia infusa. Conviértelo en hábito y así no te entrará el pánico cuando veas un problema nuevo. Por lo menos sabrás dividirlo en partes más pequeñas, lo cual siempre ayuda.

Hacer un dibujo

Seguro que has oído el dicho "Una imagen vale más que mil palabras". Como profesora de estadística, yo siempre digo que una imagen vale más que mil puntos (o al menos la mitad de los puntos del problema). Si la información que te dan o la pregunta planteada pueden expresarse de forma gráfica, debes hacerlo. Éstas son las razones:

- ✔ **Un dibujo puede ayudarte a entender de qué va el problema.** Por ejemplo, si sabes que las notas de examen tienen una distribución normal con una media de 75 y una desviación estándar de 5 (en el capítulo 9 encontrarás más información sobre la distribución normal), dibujas una campana de Gauss y marcas la media en el centro y tres desviaciones estándares por cada lado. Ahora puedes visualizar la situación planteada.
- ✔ **Un dibujo puede ayudarte a saber lo que debes encontrar.** Por ejemplo, si tienes que averiguar la probabilidad de que Roberto haya obtenido más de 70 puntos en el examen, sombreas el área que queda a la derecha del 70 en tu dibujo y ya puedes ponerte manos a la obra.
- ✔ **Tu profesor puede ver que entiendes el planteamiento básico del problema, con lo que es más probable que te cuente al menos una parte de los puntos aunque lo hagas mal.** Por otro lado, si te equivocas en la respuesta y el profesor sabe que podrías haberlo evitado haciendo un simple dibujo, no esperes que sea comprensivo contigo.
- ✔ **Los alumnos que hacen dibujos generalmente resuelven los problemas de manera correcta.** Sin un dibujo es fácil que se te vaya de la cabeza lo que te están pidiendo y cometas errores como calcular $P(X<70)$ en lugar de $P(X>70)$, por ejemplo. Además, si tienes dibujos que mirar resulta más fácil revisar los problemas y corregir posibles errores antes de entregar el examen.

A lo mejor te parece que hacer un dibujo es una pérdida de tiempo en mitad de un examen, pero lo cierto es que ahorra tiempo porque te encauza en la dirección correcta, te mantiene centrado en tus razonamientos y te ayuda a comprobar si realmente estás respondiendo a lo que te preguntan. Además, puede ayudarte a analizar la respuesta numérica final y confirmar que es correcta, o bien a detectar y corregir rápidamente un error que podría suponerte algunos puntos menos. (Te aconsejo que hagas dibujos cuando estudies, para que luego en el examen te salgan de forma natural.)

Establecer la conexión y resolver el problema

Una vez que tengas claro lo que te piden, hayas etiquetado toda la información y hayas hecho un dibujo, es hora de resolver el problema. Después del trabajo de preparación, nueve veces de cada diez recordarás un método aprendido en clase, una fórmula que contenga los elementos etiquetados o un ejemplo que hayas estudiado. Utiliza o recuerda la tabla procedimental y ponte manos a la obra. (Más información en "Elaborar una tabla procedimental".)

Dividir un problema en partes significa tener menos que pensar en cada uno de los pasos, lo cual es una gran ventaja en una situación tan estresante como un examen, en la que puedes olvidar hasta cómo te llamas (esta estrategia me recuerda un viejo chiste: "¿Cómo te comerías un elefante? Mordisco a mordisco").

En el ejemplo anterior en que nos basábamos en la superficie de una casa para predecir su precio (relee el apartado "Etiquetar la información que te dan"), conoces la media y la desviación estándar de la superficie, la media y la desviación estándar del precio, y la correlación entre ambos, y además lo tienes todo etiquetado. Te piden que encuentres la ecuación de la recta de regresión para predecir el precio de la casa a partir de su superficie, y sabes que para eso hay que resolver la ecuación $y = a + bx$, en la que x = superficie (pies cuadrados), y = precio (miles de euros), b es la pendiente de la recta de regresión, y a es la ordenada en el origen.

Ahora ya sabes lo que tienes que hacer: encontrar a y b. Como recordarás (y si no, puedes consultarlo), las fórmulas son $b = r\frac{s_y}{s_x}$ y $a = \bar{y} - b\bar{x}$. Coges los números que has etiquetado ($\bar{x}$ = 92; s_x = 16,61; $\bar{y}$ = 219,1; s_y = 2,85; y r = 0,9), los metes en las fórmulas y resuelves (casi parece una receta de cocina, ¿verdad?). La pendiente es $b = 0{,}90\frac{2{,}85}{16{,}61}$ = 0,154 y la ordenada en el origen es a=219,1 – 0,154 (92) = –204,93, de manera que la ecuación de la recta de regresión es y = –204,93 + 0,154x. (En el capítulo 18 explico la regresión con todo detalle.)

Haz las cuentas... dos veces

Todavía recuerdo los problemas que tuve con el álgebra en mi época de instituto. Para mí, durante mucho tiempo, 3×2 era igual a 5. Este error, y otros similares, me hicieron perder un montón de puntos en los exámenes, y no había manera de arreglarlo. Un día decidí que ya no iba a perder más puntos por algo tan absurdo. A partir de aquel momento escribí todos los pasos y resistí la tentación de hacer cálculos de cabeza. Cuando llegaba a un resultado final, en lugar de pasar al siguiente problema, volvía al principio y lo repasaba todo paso por paso, convencida de que probablemente habría un error en algún sitio y tenía que encontrarlo antes que el profe.

Esta estrategia me obligaba a revisar cada paso sin dar nada por sentado, como si estuviera calificando el examen de otra persona. Encontraba más errores porque nunca me saltaba un paso sin comprobarlo. De tantas veces que me corregí a mí misma, al final dejé de pensar que 3×2 eran 5. Comencé a sacar mejores notas en los exámenes, simplemente porque empecé a revisarlo todo con más atención. Eso me recuerda un viejo dicho: "El buen carpintero mide dos veces y corta una". De este modo se desperdicia mucha menos madera.

Cada vez que detectas y corriges un error antes de entregar el examen, estás ganando un montón de puntos. Encuentra tus fallos antes de que lo haga el profesor y te sorprenderás de cuánto mejorarán tus notas. No obstante, ten presente que en los exámenes tienes un tiempo limitado, de manera que conviene hacer los problemas bien al primer intento. Etiquetar toda la información, hacer dibujos y escribir las fórmulas son cosas que ayudan, y mucho.

Analizar tus respuestas

Un estadístico muy destacado que conozco tiene un papel enmarcado en su despacho. Es una página de un examen que hizo cuando estaba en el instituto. Una de las respuestas, que es un número dos, está marcada con un gran círculo rojo. ¿Por qué estaba tan mal poner un dos como respuesta? Porque el problema le pedía que encontrara una probabilidad, y las probabilidades siempre están entre 0 y 1. En consecuencia, no le contaron ningún punto por ese problema. De hecho, estoy segura de que su profesor incluso hubiera querido restarle puntos por haber cometido ese fallo (molesta mucho ver errores tan garrafales).

Comprueba siempre si tu respuesta final tiene sentido. Una desviación estándar negativa, una probabilidad mayor que 1 o una correlación de –121,23 son cosas que no van a sentar nada bien a tu profesor, y no las verá como simples errores matemáticos. Pensará que no sabes (o quizá no te importa) cómo debe ser el resultado final y que tienes enormes lagunas en los conocimientos básicos.

Si sabes que tu respuesta no puede estar bien de ninguna manera pero tampoco tienes idea de dónde te has equivocado, no pierdas más tiempo. Simplemente escribe una nota al margen diciendo que eres consciente de que la respuesta es incorrecta, pero no sabes dónde está el fallo. Así te distingues de ese otro alumno que encuentra una probabilidad de 10.524,31 (sí, he sido testigo) y se queda tan ancho.

Por cierto, quizá te estés preguntando por qué ese estadístico de talla mundial conserva esa hoja de examen enmarcada en una pared de su despacho. Dice que es para no perder la humildad. Aprende de su ejemplo y nunca pases al siguiente problema sin preguntarte antes: "¿Tiene sentido esta respuesta?".

Apéndice

Tablas de referencia

Este apéndice incluye tablas para encontrar probabilidades o valores críticos para las tres distribuciones explicadas en este libro: la distri bución *Z* (distribución normal estándar), la distribución *t* y la distribución binomial.

La tabla *Z*

La tabla A1 muestra probabilidades "menor o igual que" para la distribu ción *Z*; es decir, $p(Z \leq z)$ para un valor *z* dado. (En el capítulo 9 explico la forma de calcular valores *z* para una distribución normal; en el capítulo 11 explico la forma de calcular valores *z* para una distribución muestral.) La tabla A1 se utiliza de la manera siguiente

1. **Determina el valor *z* para tu problema concreto.**

   El valor *z* debe tener un dígito antes de la coma decimal (positivo, negativo o cero) y dos dígitos después de la coma decimal; por ejem plo *z* = 1,28, –2,69 o 0,13.

2. **Busca en la tabla la fila correspondiente al dígito de la parte entera y al primer dígito tras la coma decimal.**

   Por ejemplo, si el valor *z* es 1,28, busca la fila "1,2"; si *z* = –1,28, busca la fila "–1,2".

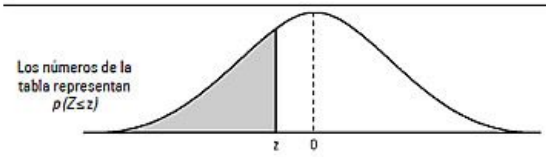
3. **Busca la columna correspondiente al segundo dígito tras la coma decimal.**

   Por ejemplo, si el valor *z* es 1,28 o –1,28, busca la columna "0,08".

4. **Encuentra la intersección de la fila y la columna de los pasos 2 y 3.**

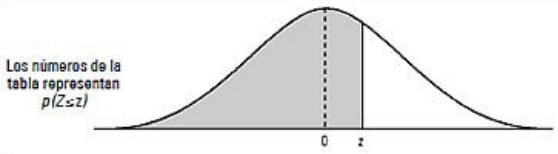
   Este número es la probabilidad de que *Z* sea menor o igual que tu valor *z*. Dicho de otro modo, has encontrado $p(Z \leq z)$. Por ejemplo, si *z* = 1,28, ves que $p(Z \leq 1{,}28) = 0{,}8997$. Para *z* = –1,28, ves que $p(Z \leq -1{,}28) = 0{,}1003$.

Tabla A-1 *(continúa)* **La tabla *Z***



z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
–3,6	0,0002	0,0002	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
–3,5	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002
–3,4	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002	0,0002
–3,3	0,0005	0,0005	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003	0,0003
–3,2	0,0007	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005
–3,1	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
–3	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
–2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
–2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
–2,1	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
–2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
–2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
–2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
–2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
–2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
–2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
–2	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
–1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
–1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
–1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
–1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
–1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
–1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
–1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
–1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
–1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
–1	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
–0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
–0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
–0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
–0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
–0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
–0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
–0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
–0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
–0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641

Tabla A-1 *(continúa)* **La tabla *Z***



z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
12	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,6	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999

La tabla *t*

La tabla A-2 muestra probabilidades de cola derecha para algunas distribuciones *t*. (En el capítulo 10 encontrarás más información sobre la distribución *t*.)

Para encontrar probabilidades de cola derecha y valores *p* para contrastes de hipótesis en los que utilices la distribución *t* (ver el capítulo 15), sigue los pasos siguientes:

1. **Determina el valor *t* cuya probabilidad de cola derecha quieres encontrar (lo llamaremos *t*) y determina el tamaño muestral (por ejemplo, *n*).**

2. **Busca la fila correspondiente a los grados de libertad *(gl)* para tu problema (por ejemplo, *n*–1). Sigue esa fila hasta encontrar los dos valores *t* entre los cuales se encuentra tu *t*.**

   Por ejemplo, si *t* es 1,60 y *n* es 7, busca en la fila que corresponde a $df = 7 - 1 = 6$. Si sigues esa fila, ves que *t* se encuentra entre los valores *t* 1,44 y 1,94.

3. **Ve a la cabecera de las columnas donde se encuentran los dos valores *t* del paso 2.**

   La probabilidad de cola derecha (mayor que) para tu valor *t* está comprendida entre los dos valores que figuran en las cabeceras de esas columnas. Por ejemplo, $t = 1,60$ se encuentra entre los valores *t* 1,44 y 1,94 ($gl = 6$); por tanto, la probabilidad de cola derecha para tu *t* está comprendida entre 0,1 (cabecera de columna para $t = 1,44$) y 0,05 (cabecera de columna para $t = 1,94$).

La penúltima fila de la tabla, donde hay una *Z* en la columna *gl*, muestra probabilidades de cola derecha (mayor que) de la distribución *Z*. (En el capítulo 10 explico la relación entre *Z* y *t*.)

Para encontrar valores t^* (valores críticos) para un intervalo de confianza cuando utilizas la distribución *t* (ver el capítulo 13), sigue los pasos siguientes:

1. **Determina qué nivel de confianza necesitas (como porcentaje).**

2. **Determina el tamaño muestral (por ejemplo, *n*).**

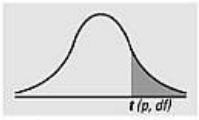
3. **En la última fila de la tabla, busca el porcentaje correspondiente a tu nivel de confianza.**

4. **Busca la intersección entre esa columna y la fila correspondiente a tus grados de libertad *(gl)*.** Ése es el valor *t* que necesitas para tu intervalo de confianza.

   Por ejemplo, para un intervalo de confianza del 95% con $gl = 6$, vemos que $t^* = 2{,}45$ (busca el 95% en la última fila y sube hasta la fila 6).

Tabla A-2 La tabla *t*

Los números de cada fila son valores de una distribución *t* con *(gl)* grados de libertad para diversas probabilidades de cola derecha (mayor que) *(p)*.



df/p	0,40	0,25	0,10	0,05	0,025	0,01	0,005	0,0005
1	0,324920	10,000000	30,077684	6,313752	12,70620	31,82052	63,65674	636,6192
2	0,288675	0,816497	1,885618	2,919986	4,30265	6,96456	9,92484	31,5991
3	0,276671	0,764892	1,637744	2,353363	3,18245	4,54070	5,84091	12,9240
4	0270722	0,740697	1,533206	2,131847	2,77645	3,74695	4,60409	8,6103
5	0,267181	0,726687	1,475884	20,015048	2,57058	3,36493	40,03214	6,8688
6	0,264835	0,717558	1,439756	1,943180	2,44691	3,14267	3,70743	5,9588
7	0,263167	0,711142	1,414924	1,894579	2,36462	2,99795	3,49948	5,4079
8	0,261921	0,706387	1,396815	1,859548	2,30600	2,89646	3,35539	50,0413
9	0,260955	0,702722	1,383029	1,833113	2,26216	2,82144	3,24984	4,7809
10	0260185	0,699812	1,372184	1,812461	2,22814	2,76377	3,16927	4,5869
11	0259556	0,697445	1,363430	1,795885	2,20099	2,71808	3,10581	4,4370
12	0259033	0,695483	1,356217	1,782288	2,17881	2,68100	30,05454	43178
13	0,258591	0,693829	1,350171	1,770933	2,16037	2,65031	30,01228	4,2208
14	0,258213	0,692417	1,345030	1,761310	2,14479	2,62449	2,97684	4,1405
15	0,257885	0,691197	1,340606	1,753050	2,13145	2,60248	2,94671	40,0728
16	0257599	0,690132	1,336757	1,745884	2,11991	2,58349	2,92078	40,0150
17	0,257347	0,689195	1,333379	1,739607	2,10982	2,56693	2,89823	3,9651
18	0,257123	0,688364	1,330391	1,734064	2,10092	2,55238	2,87844	3,9216
19	0,256923	0,687621	1,327728	1,729133	20,09302	2,53948	2,86093	3,8834
20	0,256743	0,686954	1,325341	1,724718	20,08596	2,52798	2,84534	3,8495
21	0,256580	0,686352	1,323188	1,720743	20,07961	2,51765	2,83136	3,8193
22	0256432	0,685805	1,321237	1,717144	20,07387	2,50832	2,81876	3,7921
23	0256297	0,685306	1,319460	1,713872	20,06866	2,49987	2,80734	3,7676
24	0,256173	0,684850	1,317836	1,710882	20,06390	2,49216	2,79694	3,7454
25	0,256060	0,684430	1,316345	1,708141	20,05954	2,48511	2,78744	3,7251
26	0,255955	0,684043	1,314972	1,705618	20,05553	2,47863	2,77871	3,7066
27	0,255858	0,683685	1,313703	1,703288	20,05183	2,47266	2,77068	3,6896
28	0,255768	0,683353	1,312527	1,701131	20,04841	2,46714	2,76326	3,6739
29	0,255684	0,683044	1,311434	1,699127	20,04523	2,46202	2,75639	3,6594
30	0,255605	0,682756	1,310415	1,697261	20,04227	2,45726	2,75000	3,6460
z	0,253347	0,674490	1,281552	1,644854	1,95996	2,32635	2,57583	3,2905
CI	—	—	80%	90%	95%	98%	99%	99,9%

La tabla binomial

La tabla A3 muestra probabilidades para la distribución binomial. (Ver el capítulo 8.)

Esta tabla se utiliza de la manera siguiente:

1. **Determina estos tres números para tu problema concreto:**
   - El tamaño muestral, *n*.
   - La probabilidad de éxito, *p*.
   - El valor *x* para el cual quieres conocer $p(X = x)$.

2. **Busca el apartado de la tabla A-3 que corresponde a tu *n*.**

3. **Busca la fila correspondiente a tu valor *x* y la columna correspondiente a tu *p*.**

4. **Busca la intersección entre esa fila y esa columna.** Has encontrado $p(X = x)$.

5. **Para encontrar la probabilidad de que *X* sea menor que un valor, mayor que un valor, mayor o igual que un valor, menor o igual que un valor, o esté entre dos valores, tienes que sumar los valores correspondientes de la tabla A-3 siguiendo los pasos descritos en el capítulo 8.**

Por ejemplo, si $n = 10$ y $p = 0{,}6$ y quieres saber $p(X = 9)$, vas a la sección $n = 10$, buscas la fila $x = 9$ y la columna $p = 0{,}6$ y ves que la probabilidad es 0,04.

Tabla A-3 La tabla binominal

Los números de la tabla representan $p(X = x)$ para una distribución binominal con n intentos y una probabilidad de éxito p.

Probabilidades binomiales: $\binom{n}{x} p^x (1-p)^{n-x}$

		p										
n	x	0,1	0,2	0,25	0,3	0,4	0,5	0,6	0,7	0,75	0,8	0,9
1	0	0,900	0,800	0,750	0,700	0,600	0,500	0,400	0,300	0,250	0,200	0,100
	1	0,100	0,200	0,250	0,300	0,400	0,500	0,600	0,700	0,750	0,800	0,900
2	0	0,810	0,640	0,563	0,490	0,360	0,250	0,160	0,090	0,063	0,040	0,010
	1	0,180	0,320	0,375	0,420	0,480	0,500	0,480	0,420	0,375	0,320	0,180
	2	0,010	0,040	0,063	0,090	0,160	0,250	0,360	0,490	0,563	0,640	0,810
3	0	0,729	0,512	0,422	0,343	0,216	0,125	0,064	0,027	0,016	0,008	0,001
	1	0,243	0,384	0,422	0,441	0,432	0,375	0,288	0,189	0,141	0,096	0,027
	2	0,027	0,096	0,141	0,189	0,288	0,375	0,432	0,441	0,422	0,384	0,243
	3	0,001	0,008	0,016	0,027	0,064	0,125	0,216	0,343	0,422	0,512	0,729
4	0	0,656	0,410	0,316	0,240	0,130	0,063	0,026	0,008	0,004	0,002	0,000
	1	0,292	0,410	0,422	0,412	0,346	0,250	0,154	0,076	0,047	0,026	0,004
	2	0,049	0,154	0,211	0,265	0,346	0,375	0,346	0,265	0,211	0,154	0,049
	3	0,004	0,026	0,047	0,076	0,154	0,250	0,346	0,412	0,422	0,410	0,292
	4	0,000	0,002	0,004	0,008	0,026	0,063	0,130	0,240	0,316	0,410	0,656
5	0	0,590	0,328	0,237	0,168	0,078	0,031	0,010	0,002	0,001	0,000	0,000
	1	0,328	0,410	0,396	0,360	0,259	0,156	0,077	0,028	0,015	0,006	0,000
	2	0,073	0,205	0,264	0,309	0,346	0,312	0,230	0,132	0,088	0,051	0,008
	3	0,008	0,051	0,088	0,132	0,230	0,312	0,346	0,309	0,264	0,205	0,073
	4	0,000	0,006	0,015	0,028	0,077	0,156	0,259	0,360	0,396	0,410	0,328
	5	0,000	0,000	0,001	0,002	0,010	0,031	0,078	0,168	0,237	0,328	0,590
6	0	0,531	0,262	0,178	0,118	0,047	0,016	0,004	0,001	0,000	0,000	0,000
	1	0,354	0,393	0,356	0,303	0,187	0,094	0,037	0,010	0,004	0,002	0,000
	2	0,098	0,246	0,297	0,324	0,311	0,234	0,138	0,060	0,033	0,015	0,001
	3	0,015	0,082	0,132	0,185	0,276	0,313	0,276	0,185	0,132	0,082	0,015
	4	0,001	0,015	0,033	0,060	0,138	0,234	0,311	0,324	0,297	0,246	0,098
	5	0,000	0,002	0,004	0,010	0,037	0,094	0,187	0,303	0,356	0,393	0,354
	6	0,000	0,000	0,000	0,001	0,004	0,016	0,047	0,118	0,178	0,262	0,531
7	0	0,478	0,210	0,133	0,082	0,028	0,008	0,002	0,000	0,000	0,000	0,000
	1	0,372	0,367	0,311	0,247	0,131	0,055	0,017	0,004	0,001	0,000	0,000
	2	0,124	0,275	0,311	0,318	0,261	0,164	0,077	0,025	0,012	0,004	0,000
	3	0,023	0,115	0,173	0,227	0,290	0,273	0,194	0,097	0,058	0,029	0,003
	4	0,003	0,029	0,058	0,097	0,194	0,273	0,290	0,227	0,173	0,115	0,023
	5	0,000	0,004	0,012	0,025	0,077	0,164	0,261	0,318	0,311	0,275	0,124
	6	0,000	0,000	0,001	0,004	0,017	0,055	0,131	0,247	0,311	0,367	0,372
	7	0,000	0,000	0,000	0,000	0,002	0,008	0,028	0,082	0,133	0,210	0,478

Tabla A-3 *(continúa)*

Los números de la tabla representan $p(X = x)$ para una distribución binominal con n intentos y una probabilidad de éxito p.

Probabilidades binomiales: $\binom{n}{x} p^x (1-p)^{n-x}$

		p										
n	x	0,1	0,2	0,25	0,3	0,4	0,5	0,6	0,7	0,75	0,8	0,9
8	0	0,430	0,168	0,100	0,058	0,017	0,004	0,001	0,000	0,000	0,000	0,000
	1	0,383	0,336	0,267	0,198	0,090	0,031	0,008	0,001	0,000	0,000	0,000
	2	0,149	0,294	0,311	0,296	0,209	0,109	0,041	0,010	0,004	0,001	0,000
	3	0,033	0,147	0,208	0,254	0,279	0,219	0,124	0,047	0,023	0,009	0,000
	4	0,005	0,046	0,087	0,136	0,232	0,273	0,232	0,136	0,087	0,046	0,005
	5	0,000	0,009	0,023	0,047	0,124	0,219	0,279	0,254	0,208	0,147	0,033
	6	0,000	0,001	0,004	0,010	0,041	0,109	0,209	0,296	0,311	0,294	0,149
	7	0,000	0,000	0,000	0,001	0,008	0,031	0,090	0,198	0,267	0,336	0,383
	8	0,000	0,000	0,000	0,000	0,001	0,004	0,017	0,058	0,100	0,168	0,430
9	0	0,387	0,134	0,075	0,040	0,010	0,002	0,000	0,000	0,000	0,000	0,000
	1	0,387	0,302	0,225	0,156	0,060	0,018	0,004	0,000	0,000	0,000	0,000
	2	0,172	0,302	0,300	0,267	0,161	0,070	0,021	0,004	0,001	0,000	0,000
	3	0,045	0,176	0,234	0,267	0,251	0,164	0,074	0,021	0,009	0,003	0,000
	4	0,007	0,066	0,117	0,172	0,251	0,246	0,167	0,074	0,039	0,017	0,001
	5	0,001	0,017	0,039	0,074	0,167	0,246	0,251	0,172	0,117	0,066	0,007
	6	0,000	0,003	0,009	0,021	0,074	0,164	0,251	0,267	0,234	0,176	0,045
	7	0,000	0,000	0,001	0,004	0,021	0,070	0,161	0,267	0,300	0,302	0,172
	8	0,000	0,000	0,000	0,000	0,004	0,018	0,060	0,156	0,225	0,302	0,387
	9	0,000	0,000	0,000	0,000	0,000	0,002	0,010	0,040	0,075	0,134	0,387
10	0	0,349	0,107	0,056	0,028	0,006	0,001	0,000	0,000	0,000	0,000	0,000
	1	0,387	0,268	0,188	0,121	0,040	0,010	0,002	0,000	0,000	0,000	0,000
	2	0,194	0,302	0,282	0,233	0,121	0,044	0,011	0,001	0,000	0,000	0,000
	3	0,057	0,201	0,250	0,267	0,215	0,117	0,042	0,009	0,003	0,001	0,000
	4	0,011	0,088	0,146	0,200	0,251	0,205	0,111	0,037	0,016	0,006	0,000
	5	0,001	0,026	0,058	0,103	0,201	0,246	0,201	0,103	0,058	0,026	0,001
	6	0,000	0,006	0,016	0,037	0,111	0,205	0,251	0,200	0,146	0,088	0,011
	7	0,000	0,001	0,003	0,009	0,042	0,117	0,215	0,267	0,250	0,201	0,057
	8	0,000	0,000	0,000	0,001	0,011	0,044	0,121	0,233	0,282	0,302	0,194
	9	0,000	0,000	0,000	0,000	0,002	0,010	0,040	0,121	0,188	0,268	0,387
	10	0,000	0,000	0,000	0,000	0,000	0,001	0,006	0,028	0,056	0,107	0,349
11	0	0,314	0,086	0,042	0,020	0,004	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,384	0,236	0,155	0,093	0,027	0,005	0,001	0,000	0,000	0,000	0,000
	2	0,213	0,295	0,258	0,200	0,089	0,027	0,005	0,001	0,000	0,000	0,000
	3	0,071	0,221	0,258	0,257	0,177	0,081	0,023	0,004	0,001	0,000	0,000
	4	0,016	0,111	0,172	0,220	0,236	0,161	0,070	0,017	0,006	0,002	0,000
	5	0,002	0,039	0,080	0,132	0,221	0,226	0,147	0,057	0,027	0,010	0,000
	6	0,000	0,010	0,027	0,057	0,147	0,226	0,221	0,132	0,080	0,039	0,002
	7	0,000	0,002	0,006	0,017	0,070	0,161	0,236	0,220	0,172	0,111	0,016
	8	0,000	0,000	0,001	0,004	0,023	0,081	0,177	0,257	0,258	0,221	0,071
	9	0,000	0,000	0,000	0,001	0,005	0,027	0,089	0,200	0,258	0,295	0,213
	10	0,000	0,000	0,000	0,000	0,001	0,005	0,027	0,093	0,155	0,236	0,384
	11	0,000	0,000	0,000	0,000	0,000	0,000	0,004	0,020	0,042	0,086	0,314

Tabla A-3 *(continúa)*

Los números de la tabla representan $p(X = x)$ para una distribución binominal con n intentos y una probabilidad de éxito p.

Probabilidades binomiales: $\binom{n}{x} p^x (1-p)^{n-x}$

		p										
n	x	0,1	0,2	0,25	0,3	0,4	0,5	0,6	0,7	0,75	0,8	0,9
12	0	0,282	0,069	0,032	0,014	0,002	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,377	0,206	0,127	0,071	0,017	0,003	0,000	0,000	0,000	0,000	0,000
	2	0,230	0,283	0,232	0,168	0,064	0,016	0,002	0,000	0,000	0,000	0,000
	3	0,085	0,236	0,258	0,240	0,142	0,054	0,012	0,001	0,000	0,000	0,000
	4	0,021	0,133	0,194	0,231	0,213	0,121	0,042	0,008	0,002	0,001	0,000
	5	0,004	0,053	0,103	0,158	0,227	0,193	0,101	0,029	0,011	0,003	0,000
	6	0,000	0,016	0,040	0,079	0,177	0,226	0,177	0,079	0,040	0,016	0,000
	7	0,000	0,003	0,011	0,029	0,101	0,193	0,227	0,158	0,103	0,053	0,004
	8	0,000	0,001	0,002	0,008	0,042	0,121	0,213	0,231	0,194	0,133	0,021
	9	0,000	0,000	0,000	0,001	0,012	0,054	0,142	0,240	0,258	0,236	0,085
	10	0,000	0,000	0,000	0,000	0,002	0,016	0,064	0,168	0,232	0,283	0,230
	11	0,000	0,000	0,000	0,000	0,000	0,003	0,017	0,071	0,127	0,206	0,377
	12	0,000	0,000	0,000	0,000	0,000	0,000	0,002	0,014	0,032	0,069	0,282
13	0	0,254	0,055	0,024	0,010	0,001	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,367	0,179	0,103	0,054	0,011	0,002	0,000	0,000	0,000	0,000	0,000
	2	0,245	0,268	0,206	0,139	0,045	0,010	0,001	0,000	0,000	0,000	0,000
	3	0,100	0,246	0,252	0,218	0,111	0,035	0,006	0,001	0,000	0,000	0,000
	4	0,028	0,154	0,210	0,234	0,184	0,087	0,024	0,003	0,001	0,000	0,000
	5	0,006	0,069	0,126	0,180	0,221	0,157	0,066	0,014	0,005	0,001	0,000
	6	0,001	0,023	0,056	0,103	0,197	0,209	0,131	0,044	0,019	0,006	0,000
	7	0,000	0,006	0,019	0,044	0,131	0,209	0,197	0,103	0,056	0,023	0,001
	8	0,000	0,001	0,005	0,014	0,066	0,157	0,221	0,180	0,126	0,069	0,006
	9	0,000	0,000	0,001	0,003	0,024	0,087	0,184	0,234	0,210	0,154	0,028
	10	0,000	0,000	0,000	0,001	0,006	0,035	0,111	0,218	0,252	0,246	0,100
	11	0,000	0,000	0,000	0,000	0,001	0,010	0,045	0,139	0,206	0,268	0,245
	12	0,000	0,000	0,000	0,000	0,000	0,002	0,011	0,054	0,103	0,179	0,367
	13	0,000	0,000	0,000	0,000	0,000	0,000	0,001	0,010	0,024	0,055	0,254
14	0	0,229	0,044	0,018	0,007	0,001	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,356	0,154	0,083	0,041	0,007	0,001	0,000	0,000	0,000	0,000	0,000
	2	0,257	0,250	0,180	0,113	0,032	0,006	0,001	0,000	0,000	0,000	0,000
	3	0,114	0,250	0,240	0,194	0,085	0,022	0,003	0,000	0,000	0,000	0,000
	4	0,035	0,172	0,220	0,229	0,155	0,061	0,014	0,001	0,000	0,000	0,000
	5	0,008	0,086	0,147	0,196	0,207	0,122	0,041	0,007	0,002	0,000	0,000
	6	0,001	0,032	0,073	0,126	0,207	0,183	0,092	0,023	0,008	0,002	0,000
	7	0,000	0,009	0,028	0,062	0,157	0,209	0,157	0,062	0,028	0,009	0,000
	8	0,000	0,002	0,008	0,023	0,092	0,183	0,207	0,126	0,073	0,032	0,001
	9	0,000	0,000	0,002	0,007	0,041	0,122	0,207	0,196	0,147	0,086	0,008
	10	0,000	0,000	0,000	0,001	0,014	0,061	0,155	0,229	0,220	0,172	0,035
	11	0,000	0,000	0,000	0,000	0,003	0,022	0,085	0,194	0,240	0,250	0,114
	12	0,000	0,000	0,000	0,000	0,001	0,006	0,032	0,113	0,180	0,250	0,257
	13	0,000	0,000	0,000	0,000	0,000	0,001	0,007	0,041	0,083	0,154	0,356
	14	0,000	0,000	0,000	0,000	0,000	0,000	0,001	0,007	0,018	0,044	0,229

Tabla A-3 *(continúa)*

Los números de la tabla representan $p(X = x)$ para una distribución binominal con n intentos y una probabilidad de éxito p.

Probabilidades binomiales: $\binom{n}{x} p^x (1-p)^{n-x}$

		p										
n	x	0,1	0,2	0,25	0,3	0,4	0,5	0,6	0,7	0,75	0,8	0,9
15	0	0,206	0,035	0,013	0,005	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,343	0,132	0,067	0,031	0,005	0,000	0,000	0,000	0,000	0,000	0,000
	2	0,267	0,231	0,156	0,092	0,022	0,003	0,000	0,000	0,000	0,000	0,000
	3	0,129	0,250	0,225	0,170	0,063	0,014	0,002	0,000	0,000	0,000	0,000
	4	0,043	0,188	0,225	0,219	0,127	0,042	0,007	0,001	0,000	0,000	0,000
	5	0,010	0,103	0,165	0,206	0,186	0,092	0,024	0,003	0,001	0,000	0,000
	6	0,002	0,043	0,092	0,147	0,207	0,153	0,061	0,012	0,003	0,001	0,000
	7	0,000	0,014	0,039	0,081	0,177	0,196	0,118	0,035	0,013	0,003	0,000
	8	0,000	0,003	0,013	0,035	0,118	0,196	0,177	0,081	0,039	0,014	0,000
	9	0,000	0,001	0,003	0,012	0,061	0,153	0,207	0,147	0,092	0,043	0,002
	10	0,000	0,000	0,001	0,003	0,024	0,092	0,186	0,206	0,165	0,103	0,010
	11	0,000	0,000	0,000	0,001	0,007	0,042	0,127	0,219	0,225	0,188	0,043
	12	0,000	0,000	0,000	0,000	0,002	0,014	0,063	0,170	0,225	0,250	0,129
	13	0,000	0,000	0,000	0,000	0,000	0,003	0,022	0,092	0,156	0,231	0,267
	14	0,000	0,000	0,000	0,000	0,000	0,000	0,005	0,031	0,067	0,132	0,343
	15	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,005	0,013	0,035	0,206
20	0	0,122	0,012	0,003	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,270	0,058	0,021	0,007	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	2	0,285	0,137	0,067	0,028	0,003	0,000	0,000	0,000	0,000	0,000	0,000
	3	0,190	0,205	0,134	0,072	0,012	0,001	0,000	0,000	0,000	0,000	0,000
	4	0,090	0,218	0,190	0,130	0,035	0,005	0,000	0,000	0,000	0,000	0,000
	5	0,032	0,175	0,202	0,179	0,075	0,015	0,001	0,000	0,000	0,000	0,000
	6	0,009	0,109	0,169	0,192	0,124	0,037	0,005	0,000	0,000	0,000	0,000
	7	0,002	0,055	0,112	0,164	0,166	0,074	0,015	0,001	0,000	0,000	0,000
	8	0,000	0,022	0,061	0,114	0,180	0,120	0,035	0,004	0,001	0,000	0,000
	9	0,000	0,007	0,027	0,065	0,160	0,160	0,071	0,012	0,003	0,000	0,000
	10	0,000	0,002	0,010	0,031	0,117	0,176	0,117	0,031	0,010	0,002	0,000
	11	0,000	0,000	0,003	0,012	0,071	0,160	0,160	0,065	0,027	0,007	0,007
	12	0,000	0,000	0,001	0,004	0,035	0,120	0,180	0,114	0,061	0,022	0,000
	13	0,000	0,000	0,000	0,001	0,015	0,074	0,166	0,164	0,112	0,055	0,002
	14	0,000	0,000	0,000	0,000	0,005	0,037	0,124	0,192	0,169	0,109	0,009
	15	0,000	0,000	0,000	0,000	0,001	0,015	0,075	0,179	0,202	0,175	0,032
	16	0,000	0,000	0,000	0,000	0,000	0,005	0,035	0,130	0,190	0,218	0,090
	17	0,000	0,000	0,000	0,000	0,000	0,001	0,012	0,072	0,134	0,205	0,190
	18	0,000	0,000	0,000	0,000	0,000	0,000	0,003	0,028	0,067	0,137	0,285
	19	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,007	0,021	0,058	0,270
	20	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,001	0,003	0,012	0,122

Estadística para Dummies
Deborah J. Rumsey
ISBN edición en papel: 978-84-329-0157-7

Título original: *Statistics for Dummies, 2rd edition*

Grupo Planeta
Av. Diagonal, 662-664
08034 Barcelona (España)
www.planetadelibros.com

Edición publicada mediante acuerdo con Wiley Publishing, Inc.

Primera edición en libro electrónico (epub): septiembre 2013

ISBN: 978-84-329-0194-2 (epub)

Conversión a libro electrónico: Newcomlab, S. L. L.
www.newcomlab.com

www.paradummies.es