David Diez Científico de Datos OpenIntro

Mine Çetinkaya-Rundel Profesora Asociada de la Práctica, Universidad de Duke Educadora Profesional, RStudio

Christopher D Barr Analista de Inversiones Varadero Capital

2019. Cuarta Edición. Actualizado: 30 de diciembre de 2024.

Versión traducida automatizadamente mediante el uso de un "Large Language Model" en marzo de 2025. Algunas palabras y términos pueden no ser los más apropiados dentro del contexto estadístico. Prontamente se realizará una revisión preliminar junto a las correcciones

Este libro se puede descargar como un PDF gratuito en openintro.org/book/os. Este libro de texto también está disponible bajo una licencia Creative Commons, con los archivos fuente alojados en Github.

Tabla de Contenidos

Prefacio

OpenIntro Statistics cubre un primer curso de estadística, proporcionando una introducción rigurosa a la estadística aplicada que es clara, concisa y accesible. Este libro fue escrito teniendo en mente el nivel de pregrado, pero también es popular en escuelas secundarias y cursos de posgrado.

Esperamos que los lectores se lleven tres ideas de este libro, además de formar una base de pensamiento y métodos estadísticos.

- La estadística es un campo aplicado con una amplia gama de aplicaciones prácticas.
- No tienes que ser un gurú de las matemáticas para aprender de datos reales e interesantes.
- Los datos son desordenados y las herramientas estadísticas son imperfectas. Pero, cuando comprendes las fortalezas y debilidades de estas herramientas, puedes usarlas para aprender sobre el mundo.

Resumen del libro de texto

Los capítulos de este libro son los siguientes:

- 1. Introducción a los datos. Estructuras de datos, variables y técnicas básicas de recopilación de datos.
- 2. Resumen de datos. Resúmenes de datos, gráficos y un anticipo de la inferencia mediante la aleatorización.
- 3. Probabilidad. Principios básicos de probabilidad.
- 4. Distribuciones de variables aleatorias. El modelo normal y otras distribuciones clave.
- 5. Fundamentos para la inferencia. Ideas generales para la inferencia estadística en el contexto de la estimación de la proporción poblacional.

- 6. Inferencia para datos categóricos. Inferencia para proporciones y tablas utilizando las distribuciones normal y chi-cuadrado.
- 7. Inferencia para datos numéricos. Inferencia para una o dos medias muestrales utilizando la distribución t, potencia estadística para comparar dos grupos y también comparaciones de muchas medias utilizando ANOVA.
- 8. Introducción a la regresión lineal. Regresión para un resultado numérico con una variable predictora. La mayor parte de este capítulo podría cubrirse después del Capítulo 1.
- 9. Regresión múltiple y logística. Regresión para datos numéricos y categóricos utilizando muchos predictores.

OpenIntro Statistics admite flexibilidad en la elección y el orden de los temas. Si el objetivo principal es llegar a la regresión múltiple (Capítulo 9) lo más rápido posible, entonces los siguientes son los requisitos previos ideales:

- Capítulo 1, Secciones 2.1, y Sección 2.2 para una sólida introducción a las estructuras de datos y resúmenes estadísticos que se utilizan en todo el libro.
- Sección 4.1 para una sólida comprensión de la distribución normal.
- Capítulo 5 para establecer el conjunto básico de herramientas de inferencia.
- Sección 7.1 para proporcionar una base para la distribución t
- Capítulo 8 para establecer ideas y principios para la regresión de un solo predictor.

Ejemplos y ejercicios

Se proporcionan ejemplos para establecer una comprensión de cómo aplicar los métodos

EJEMPLO 0.1

Esto es un ejemplo. Cuando se hace una pregunta aquí, ¿dónde se puede encontrar la respuesta?

¡La respuesta se puede encontrar aquí, en la sección de solución del ejemplo!

Cuando creemos que el lector debería estar listo para intentar determinar la solución a un ejemplo, lo enmarcamos como Práctica Guiada.

PRÁCTICA GUIADA 0.2

El lector puede verificar o aprender la respuesta a cualquier problema de Práctica Guiada revisando la solución completa en una nota al pie.1

También se proporcionan ejercicios al final de cada sección, así como ejercicios de repaso al final de cada capítulo. Las soluciones se dan para los ejercicios impares en el Apéndice A.

Recursos adicionales

Videos explicativos, diapositivas, laboratorios de software estadístico, conjuntos de datos utilizados en el libro de texto y mucho más están disponibles en

openintro.org/os

También hemos mejorado la capacidad de acceder a los datos de este libro mediante la adición del Apéndice B, que proporciona información adicional para cada uno de los conjuntos de datos utilizados en el texto principal y es nuevo en la Cuarta Edición. También se proporcionan guías en línea para cada uno de estos conjuntos de datos en openintro.org/data y a través de un paquete complementario de R.

Agradecemos todos los comentarios, así como los informes de cualquier errata a través del sitio web. Un enlace corto para informar de una nueva errata o revisar las erratas conocidas es openintro.org/os/typos.

Para aquellos que se centran en la estadística a nivel de escuela secundaria, consideren Estadística Avanzada para la Escuela Secundaria, que es una versión de OpenIntro Statistics que ha sido ampliamente personalizada por Leah Dorazio para cursos de escuela secundaria y AP® Statistics.

Agradecimientos

Este proyecto no sería posible sin la pasión y dedicación de muchas más personas además de las que figuran en la lista de autores. Los autores desean agradecer al Personal de OpenIntro por su participación y contribuciones continuas. También estamos muy agradecidos a los cientos de estudiantes e instructores que nos han proporcionado valiosos comentarios desde que comenzamos a publicar contenido del libro en 2009.

También queremos agradecer a los muchos profesores que ayudaron a revisar esta edición, incluyendo a Laura Acion, Matthew E. Aiello-Lammens, Jonathan Akin, Stacey C.

Behrensmeyer, Juan Gomez, Jo Hardin, Nicholas Horton, Danish Khan, Peter H.M. Klaren, Jesse Mostipak, Jon C. New, Mario Orsi, Steve Phelps, y David Rockoff. Agradecemos todos sus comentarios, que nos ayudaron a ajustar el texto de manera significativa y mejoraron enormemente este libro.

1Los problemas de Práctica Guiada están destinados a expandir tu pensamiento, y puedes verificar tu respuesta revisando la solución de la nota al pie para cualquier Práctica Guiada.

Capítulo 1

7

Introducción a los datos

- **1.1 Estudio de caso: uso de stents para prevenir accidentes cerebrovasculares**
- **1.2 Conceptos básicos de datos**
- **1.3 Principios y estrategias de muestreo**
- **1.4 Experimentos**

Los científicos buscan responder preguntas utilizando métodos rigurosos y observaciones cuidadosas. Estas observaciones, recopiladas de notas de campo, encuestas y experimentos, forman la columna vertebral de una investigación estadística y se denominan datos. La estadística es el estudio de la mejor manera de recopilar, analizar y sacar conclusiones a partir de datos, y en este primer capítulo, nos centraremos tanto en las propiedades de los datos como en la recopilación de datos.

Para videos, diapositivas y otros recursos, visite www.openintro.org/os

1.1 Estudio de caso: uso de stents para prevenir accidentes cerebrovasculares

La sección 1.1 presenta un desafío clásico en estadística: evaluar la eficacia de un tratamiento médico. Los términos de esta sección, y de hecho de gran parte de este capítulo, se revisarán

más adelante en el texto. El plan por ahora es simplemente tener una idea del papel que la estadística puede desempeñar en la práctica.

En esta sección consideraremos un experimento que estudia la eficacia de los stents en el tratamiento de pacientes con riesgo de accidente cerebrovascular. Los stents son dispositivos que se colocan dentro de los vasos sanguíneos que ayudan en la recuperación del paciente después de eventos cardíacos y reducen el riesgo de un ataque cardíaco o muerte adicional. Muchos médicos esperaban que hubiera beneficios similares para los pacientes con riesgo de accidente cerebrovascular. Comenzamos escribiendo la pregunta principal que los investigadores esperan responder:

¿El uso de stents reduce el riesgo de accidente cerebrovascular?

Los investigadores que hicieron esta pregunta realizaron un experimento con 451 pacientes en riesgo. Cada paciente voluntario fue asignado aleatoriamente a uno de dos grupos:

Grupo de tratamiento. Los pacientes en el grupo de tratamiento recibieron un stent y manejo médico. El manejo médico incluyó medicamentos, manejo de factores de riesgo y ayuda en la modificación del estilo de vida.

Grupo de control. Los pacientes en el grupo de control recibieron el mismo manejo médico que el grupo de tratamiento, pero no recibieron stents.

Los investigadores asignaron aleatoriamente a 224 pacientes al grupo de tratamiento y a 227 al grupo de control. En este estudio, el grupo de control proporciona un punto de referencia contra el cual podemos medir el impacto médico de los stents en el grupo de tratamiento.

Los investigadores estudiaron el efecto de los stents en dos momentos: 30 días después de la inscripción y 365 días después de la inscripción. Los resultados de 5 pacientes se resumen en la Figura 1.1. Los resultados de los pacientes se registran como “accidente cerebrovascular” o “sin evento”, lo que representa si el paciente tuvo o no un accidente cerebrovascular al final de un período de tiempo.

Paciente	grupo	0-30 días	0-365 días
1	tratamiento	sin evento	sin evento
2	tratamiento	accidente cerebrovascular	accidente cerebrovascular
3	tratamiento	sin evento	sin evento
450	control	sin evento	sin evento
451	control	sin evento	sin evento

Figura 1.1: Resultados de cinco pacientes del estudio de stents.

Considerar los datos de cada paciente individualmente sería un camino largo y engorroso para responder la pregunta de investigación original. En cambio, realizar un análisis de datos

estadístico nos permite considerar todos los datos a la vez. La Figura 1.2 resume los datos brutos de una manera más útil. En esta tabla, podemos ver rápidamente lo que sucedió durante todo el estudio. Por ejemplo, para identificar el número de pacientes en el grupo de tratamiento que sufrieron un accidente cerebrovascular en un plazo de 30 días, miramos en el lado izquierdo de la tabla en la intersección del tratamiento y el accidente cerebrovascular: 33.

	0-30 días		0-365 días	
	accidente cerebrovascular	sin evento	accidente cerebrovascular	sin evento
tratamiento	33	191	45	179
control	13	214	28	199
Total	46	405	73	378

Figura 1.2: Estadísticas descriptivas para el estudio de stents.

De los 224 pacientes en el grupo de tratamiento, 45 sufrieron un accidente cerebrovascular al final del primer año. Usando estos dos números, calcule la proporción de pacientes en el grupo de tratamiento que sufrieron un accidente cerebrovascular al final de su primer año. (Tenga en cuenta: las respuestas a todos los ejercicios de práctica guiada se proporcionan mediante notas al pie.)1

Podemos calcular estadísticas resumidas de la tabla. Una estadística resumida es un solo número que resume una gran cantidad de datos. Por ejemplo, los resultados principales del estudio después de 1 año podrían describirse mediante dos estadísticas resumidas: la proporción de personas que tuvieron un accidente cerebrovascular en los grupos de tratamiento y control.

Proporción de personas que tuvieron un accidente cerebrovascular en el grupo de tratamiento (stent): $45/224 = 0.20 = 20\%$.

Proporción de personas que tuvieron un accidente cerebrovascular en el grupo de control: $28/227 = 0.12 = 12\%$.

Estas dos estadísticas resumidas son útiles para buscar diferencias en los grupos, y nos espera una sorpresa: ¡un 8% adicional de pacientes en el grupo de tratamiento tuvo un accidente cerebrovascular! Esto es importante por dos razones. Primero, es contrario a lo que esperaban los médicos, que era que los stents reducirían la tasa de accidentes cerebrovasculares. En segundo lugar, conduce a una pregunta estadística: ¿muestran los datos una diferencia "real" entre los grupos?

Esta segunda pregunta es sutil. Suponga que lanza una moneda 100 veces. Si bien la probabilidad de que una moneda caiga en cara en cualquier lanzamiento de moneda dado es del 50%, probablemente no observemos exactamente 50 caras. Este tipo de fluctuación es parte de casi cualquier tipo de proceso de generación de datos. Es posible que la diferencia del 8% en el estudio de stents se deba a esta variación natural. Sin embargo, cuanto mayor sea

la diferencia que observemos (para un tamaño de muestra particular), menos creíble es que la diferencia se deba al azar. Entonces, lo que realmente estamos preguntando es lo siguiente: ¿es la diferencia tan grande que deberíamos rechazar la noción de que se debió al azar?

Si bien aún no tenemos nuestras herramientas estadísticas para abordar completamente esta pregunta por nuestra cuenta, podemos comprender las conclusiones del análisis publicado: hubo evidencia convincente de daño por stents en este estudio de pacientes con accidente cerebrovascular.

Tenga cuidado: no generalice los resultados de este estudio a todos los pacientes y a todos los stents. Este estudio observó a pacientes con características muy específicas que se ofrecieron como voluntarios para ser parte de este estudio y que pueden no ser representativos de todos los pacientes con accidente cerebrovascular. Además, existen muchos tipos de stents y este estudio solo consideró el stent Wingspan autoexpandible (Boston Scientific). Sin embargo, este estudio nos deja con una lección importante: debemos mantener los ojos abiertos a las sorpresas.

1La proporción de los 224 pacientes que sufrieron un accidente cerebrovascular en un plazo de 365 días: $45/224 = 0.20$.

Fig. 1 El área apropiada

Ejercicios

del oído (que representa el nervio ciático) que es probablemente

consentimiento informado para la participación en el estudio.

antes de aplicar NCT (T0).

123

La intensidad de la migraña se midió por medio de un VAS.

En el grupo A, se eligió un algómetro específico que ejercía una presión máxima de 250 g (SEDATELEC, Francia) para identificar los puntos sensibles con la prueba de dolor-presión (PPT). Cada punto sensible ubicado dentro del área identificada por el estudio piloto (Fig. 1, área M) se probó con NCT durante 10 s comenzando desde el pabellón auricular, que era ipsilateral, hasta el lado de

con dolor de cabeza.

1.1 Migraña y acupuntura, Parte I. Una migraña es un tipo de dolor de cabeza particularmente doloroso, que los pacientes a veces desean tratar con acupuntura. Para determinar si la acupuntura alivia el dolor de la migraña, los investigadores realizaron un estudio controlado aleatorizado donde 89 mujeres diagnosticadas con migrañas fueron asignadas aleatoriamente a uno de dos grupos: tratamiento o control. 43 pacientes en el grupo de tratamiento recibieron acupuntura que está específicamente diseñada para tratar

las migrañas. 46 pacientes en el grupo de control recibieron acupuntura placebo (inserción de agujas en ubicaciones que no son puntos de acupuntura). 24 horas después de que los pacientes recibieron acupuntura, se les preguntó si estaban libres de dolor. Los resultados se resumen en la tabla de contingencia a continuación.2 S174 Neurol Sci (2011) 32 (Supl 1):S173–S175

		Pain free		
		Yes	No	Total
Group	Treatment	10	33	43
	Control	2	44	46
	Total	12	77	89

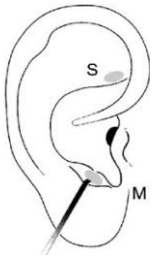


Figure from the original paper displaying the appropriate area (M) versus the inappropriate area (S) used in the treatment of migraine attacks.

- inapropiado en términos de dar un efecto terapéutico sobre los ataques de migraña, ya que no tiene correlación somatotópica En el grupo B, la rama inferior del antélix fue (a) ¿Qué porcentaje de pacientes en el grupo de tratamiento estaban libres de dolor 24 horas después de recibir acupuntura?
  - probado repetidamente con el algómetro durante unos 30 s para (b) ¿Qué porcentaje estaban libres de dolor en el grupo de control?
  - asegurarse de que no fuera sensible. En los mapas auriculares francés y chino, esta área corresponde a la representación (c) ¿En qué grupo un mayor porcentaje de pacientes se liberó del dolor 24 horas después de recibir acupuntura?
- Materiales y métodos El estudio inscribió a 94 mujeres, diagnosticadas como migraña sin aura según la Clasificación Internacional de Trastornos de Cefalea [5], que posteriormente fueron examinadas en el Centro de Cefaleas de Mujeres, Departamento de Ginecología. del nervio ciático (Fig. 1, área S) y se utiliza específicamente para tratar el dolor ciático. Se insertaron cuatro agujas en esta área, dos para cada oído. En todos los pacientes, la acupuntura auricular siempre fue realizada por un acupunturista experimentado. El análisis de los diarios que recopilaban datos de VAS fue realizado por un (d) Sus hallazgos hasta ahora podrían sugerir que la acupuntura es un tratamiento eficaz para las migrañas para todas las personas que sufren de migrañas. Sin embargo, esta no es la única conclusión posible que se puede extraer en función de sus hallazgos hasta ahora. ¿Cuál es otra posible explicación para la diferencia observada entre los porcentajes de pacientes que están libres de dolor 24 horas después de recibir acupuntura en los dos grupos?

ología y Obstetricia de la Universidad de Turín. Todas fueron incluidas en el estudio durante un ataque de migraña siempre que hubiera comenzado no más de 4 h antes. De acuerdo con una lista de aleatorización hecha por computadora predeterminada, las pacientes elegibles fueron asignadas aleatoria y ciegamente a los dos grupos siguientes: grupo A (n = 46) (edad promedio 35,93 años, rango 15–60), grupo B (n = 48) (edad promedio 33,2 años, rango 16–58). Antes de la inscripción, a cada paciente se le pidió que diera un operador imparcial que no conocía el grupo

en el que estaba cada paciente. Los valores promedio de VAS en los grupos A y B se calcularon en los diferentes momentos del estudio, y se realizó una evaluación estadística de las diferencias entre los valores obtenidos en T0, T1, T2, T3 y T4 en los dos grupos estudiados utilizando un análisis de varianza (ANOVA) para medidas repetidas seguido de la prueba t múltiple de Bonferroni para identificar la fuente de varianza. **1.2 Sinusitis y antibióticos, Parte I.** Los investigadores que estudiaron el efecto del tratamiento antibiótico para la sinusitis aguda en comparación con los tratamientos sintomáticos asignaron aleatoriamente a 166 adultos diagnosticados con sinusitis aguda a uno de dos grupos: tratamiento o control. Los participantes del estudio recibieron un curso de 10 días de amoxicilina (un antibiótico) o un placebo similar en apariencia y sabor. El placebo consistió en tratamientos sintomáticos como acetaminofeno, descongestionantes nasales, etc. Al final del período de 10 días, se preguntó a los pacientes si experimentaban una mejora en los síntomas. La distribución de las respuestas se resume a continuación.3

	Además, para evaluar la diferencia entre el grupo By el grupo A, siempre se realizó una prueba t para datos no pareados	Mejora autoinformada		
	se realizó para cada nivel de la variable ‘’tiempo’’. En el caso de			
		Sí	proporciones, se aplicó una prueba de Chi cuadrado. Todos los análisisNose realizaron utilizando el Paquete Estadístico para lo Social	Total

	Además, para evaluar la diferencia entre el grupo By el grupo A, siempre se realizó una prueba t para datos no pareados	Mejora autoinformada		
	Tratamiento	66	19Software Sciences (SPSS). Todos los valores dados en el	85
Grupo	ControlEl siguiente texto se informa como media aritmética (±SEM).	65	16	81
	Total	131	35	166

- dolor cefálico prevalente. Si la prueba fue positiva y la reducción fue de al menos el 25% con respecto a la base, una semi-Resultados (a) ¿Qué porcentaje de pacientes en el grupo de tratamiento experimentó una mejora en los síntomas?
- aguja permanente (ASP SEDATELEC, Francia) se insertó después de 1 min. Por el contrario, si el dolor no disminuía Sólo 89 pacientes de todo el grupo de 94 (43 en el grupo A, 46 en el grupo B) completaron el experimento. Cuatro pacientes (b) ¿Qué porcentaje experimentó una mejora en los síntomas en el grupo de control?
- después de 1 min, se desafió un punto sensible adicional en el se retiraron del estudio, porque experimentaron una (c) ¿En qué grupo un mayor porcentaje de pacientes experimentó una mejora en los síntomas?
- misma área y así sucesivamente. Cuando los pacientes se dieron cuenta de una disminución inicial del dolor en todas las zonas de la cabeza afectadas, fueron invitados a usar una tarjeta de diario específica para calificar la intensidad del dolor con un VAS en los siguientes intervalos: después de 10 min (T1), después de 30 min (T2), después de 60 min (T3), después de 120 min (T4) y después de 24 h (T5). exacerbación insoportable del dolor en el período anterior al último control a las 24 h (dos del grupo A y dos del grupo B) y fueron excluidos del análisis estadístico ya que solicitaron la extracción de las agujas. Una paciente del grupo A no dio su consentimiento para el implante de las agujas semipermanentes. En el grupo A, el número medio de (d) Sus hallazgos hasta

ahora podrían sugerir una diferencia real en la eficacia de los tratamientos antibióticos y placebo para mejorar los síntomas de la sinusitis. Sin embargo, esta no es la única conclusión posible que se puede extraer en función de sus hallazgos hasta ahora. ¿Cuál es otra posible explicación para la diferencia observada entre los porcentajes de pacientes en los grupos de tratamiento con antibióticos y placebo que experimentan una mejora en los síntomas de la sinusitis?

2G. Allais et al. "Acupuntura auricular en el tratamiento de los ataques de migraña: un ensayo aleatorizado sobre la eficacia de puntos de acupuntura apropiados versus inapropiados". En: Neurological Sci. 32.1 (2011), pp. 173–175.

3J.M. Garbutt et al. "Amoxicilina para la rinosinusitis aguda: un ensayo controlado aleatorizado". En: JAMA: The Journal of the American Medical Association 307.7 (2012), pp. 685–692.

1.2 Conceptos básicos de datos

La organización y descripción eficaz de los datos es un primer paso en la mayoría de los análisis. Esta sección presenta la matriz de datos para organizar los datos, así como cierta terminología sobre las diferentes formas de datos que se utilizarán a lo largo de este libro.

1.2.1 Observaciones, variables y matrices de datos

La Figura 1.3 muestra las filas 1, 2, 3 y 50 de un conjunto de datos de 50 préstamos seleccionados aleatoriamente ofrecidos a través de Lending Club, que es una empresa de préstamos entre pares. Estas observaciones se denominarán el conjunto de datos loan50.

Cada fila de la tabla representa un único préstamo. El nombre formal para una fila es caso o unidad observacional. Las columnas representan características, llamadas variables, para cada uno de los préstamos. Por ejemplo, la primera fila representa un préstamo de $22,000 con una tasa de interés del 10.90%, donde el prestatario tiene su sede en Nueva Jersey (NJ) y tiene un ingreso de $59,000.

PRÁCTICA GUIADA 1.2

¿Cuál es la calificación del primer préstamo en la Figura 1.3? ¿Y cuál es el estado de propiedad de la vivienda del prestatario para ese primer préstamo? Para estas preguntas de Práctica Guiada, puedes verificar tu respuesta en la nota al pie.4

En la práctica, es especialmente importante hacer preguntas aclaratorias para asegurar que se comprendan aspectos importantes de los datos. Por ejemplo, siempre es importante asegurarse de que sabemos lo que significa cada variable y las unidades de medida. Las descripciones de las variables loan50 se dan en la Figura 1.4.

	monto del préstamo	tasa de interés	plazo	calificación	estado	ingreso total	propiedad de vivienda
1	22000	10.90	60.00	B	NJ	59000.00	alquiler
2	6000	9.92	36.00	B	CA	60000.00	alquiler
3	25000	26.30	36.00	E	SC	75000.00	hipoteca
50	15000	6.08	36.00	A	TX	77500.00	hipoteca

Figura 1.3: Cuatro filas de la matriz de datos loan50.

variable	descripción
monto del préstamo	Monto del préstamo recibido, en dólares estadounidenses.
tasa de interés	Tasa de interés del préstamo, en un porcentaje anual.
plazo	La duración del préstamo, que siempre se establece como un número entero de meses.
calificación	Calificación del préstamo, que toma valores de A a G y representa la calidad del préstamo y su probabilidad de ser pagado.
estado	Estado de EE. UU. donde reside el prestatario.
ingreso total	Ingreso total del prestatario, incluyendo cualquier segundo ingreso, en dólares estadounidenses.
propiedad de vivienda	Indica si la persona posee, posee pero tiene una hipoteca o alquila.

Figura 1.4: Variables y sus descripciones para el conjunto de datos loan50.

Los datos en la Figura 1.3 representan una matriz de datos, que es una forma conveniente y común de organizar los datos, especialmente si se recopilan datos en una hoja de cálculo. Cada fila de una matriz de datos corresponde a un caso único (unidad de observación), y cada columna corresponde a una variable.

4La calificación del préstamo es B, y el prestatario alquila su residencia.

Al registrar datos, usa una matriz de datos a menos que tengas una muy buena razón para usar una estructura diferente. Esta estructura permite que se agreguen nuevos casos como filas o nuevas variables como nuevas columnas.

PRÁCTICA GUIADA 1.3

Las calificaciones de tareas, cuestionarios y exámenes en un curso a menudo se registran en un libro de calificaciones que toma la forma de una matriz de datos. ¿Cómo podría organizar los datos de las calificaciones utilizando una matriz de datos?5

PRÁCTICA GUIADA 1.4

Consideramos datos de 3142 condados en los Estados Unidos, que incluyen el nombre de cada condado, el estado donde reside, su población en 2017, cómo cambió su población de 2010 a 2017, la tasa de pobreza y seis características adicionales. ¿Cómo se podrían organizar estos datos en una matriz de datos?6

Los datos descritos en la Práctica Guiada 1.4 representan el conjunto de datos del condado, que se muestra como una matriz de datos en la Figura 1.5. Las variables se resumen en la Figura 1.6.

5Hay varias estrategias que se pueden seguir. Una estrategia común es que cada estudiante esté representado por una fila y luego agregar una columna para cada tarea, cuestionario o examen. Bajo esta configuración, es fácil revisar una sola línea para comprender el historial de calificaciones de un estudiante. También debe haber columnas para incluir información del estudiante, como una columna para enumerar los nombres de los estudiantes.

6Cada condado puede verse como un caso, y hay once datos registrados para cada caso. Una tabla con 3142 filas y 11 columnas podría contener estos datos, donde cada fila representa un condado y cada columna representa una pieza de información particular.

	nombre	estado	pob	cambiopob	pobreza	propiedad de vivienda	unidadmúltiple	transporte	empleo	edumediana	ingresohogarmediana
1	Autauga	Alabama	55504	1.48	13.7	77.5	7.2	3.86	sí	universidad algunos	55317
2	Baldwin	Alabama	212628	9.19	11.8	76.7	22.6	3.99	sí	universidad algunos	52562
3	Barbour	Alabama	25270	-6.22	27.2	68.0	11.1	5.90	no	diploma	33368
4	Bibb	Alabama	22668	0.73	15.2	82.9	6.6	4.39	sí	diploma	43404
5	Blount	Alabama	58013	0.68	15.6	82.0	3.7	4.02	sí	diploma	47412
6	Bullock	Alabama	10309	-2.28	28.5	76.9	9.9	4.93	no	diploma	29655

	nombre	estado	pob	cambiopob	pobreza	propiedad de vivienda	unidadmúltiple	tasadesempleo	metro	edumediana	ingresohogarmediana
7	Butler	Alabama	19825	-2.69	24.4	69.0	13.7	5.49	no	diploma	36326
8	Calhoun	Alabama	114728	-1.51	18.6	70.7	14.3	4.93	sí	universidad,algunos	43686
9	Chambers	Alabama	33713	-1.20	18.8	71.4	8.7	4.08	no	diploma	37342
10	Cherokee	Alabama	25857	-0.60	16.1	77.5	4.3	4.05	no	diploma	40041
3142	Weston	Wyoming	6927	-2.93	14.4	77.9	6.5	3.98	no	universidad,algunos	59105

Figura 1.5: Once filas del conjunto de datos del condado.

variable	descripción
nombre	nombre.Condado
estado	Columbia.deDistritoeloreside,condadoeldondeEstado
pob	2017.enPoblación
cambiopob	valorelejemplo,Por2017.a2010depoblaciónlaencambioPorcentaje 1.48%porincrementadocondadoesteparapoblaciónlasignificafilap 2017.a2010de
pobreza	pobreza.enpoblaciónladePorcentaje
propiedad de vivienda	propietario,elconviveocasapropiasuenvivequepoblaciónladePorce casa.lapropiaquienpadresconviviendoniñosp.ej.
unidadmúltiple	mentos.apartep.ej.estructuras,multi- unidadensonqueunidadesviviendodePorcentaje
tasadesempleo	porciento.acomotasaDesempleo
metro	área.metropolitanaacontienecondadoelSi
edumediana	diploma,hshs,debajoentrevaloratomarpuedequenivel,educaciónM licenciaturas.yuniversidad,algunos
ingresohogarmediana	equivaleingresohogarmayor.adonde15condado,quienlosocupante

Figura 1.6: Variables y sus descripciones para el conjunto de datos del condado.

1.2.2 Tipos de variables

Examine las variables tasa de desempleo, población, estado y educación mediana en el conjunto de datos del condado. Cada una de estas variables es inherentemente diferente de las otras tres, pero algunas comparten ciertas características.

Primero considere la tasa de desempleo, que se dice que es una variable numérica ya que puede tomar una amplia gama de valores numéricos, y es sensato sumar, restar o tomar promedios con esos valores. Por otro lado, no clasificaríamos una variable que informa los códigos de área telefónica como numérica, ya que el promedio, la suma y la diferencia de los códigos de área no tienen un significado claro.

La variable de población también es numérica, aunque parece ser un poco diferente de la tasa de desempleo. Esta variable del conteo de la población solo puede tomar números enteros no negativos (0, 1, 2, …). Por esta razón, se dice que la variable de población es discreta, ya que solo puede tomar valores numéricos con saltos. Por otro lado, se dice que la variable de la tasa de desempleo es continua.

La variable estado puede tomar hasta 51 valores después de contabilizar Washington, DC: AL, AK, …, y WY. Debido a que las respuestas en sí mismas son categorías, estado se denomina variable categórica y los valores posibles se denominan niveles de la variable.

Finalmente, considere la variable educación mediana, que describe el nivel de educación mediano de los residentes del condado y toma los valores por debajo de hs, diploma de hs, algo de universidad o licenciatura en cada condado. Esta variable parece ser un híbrido: es una variable categórica, pero los niveles tienen un ordenamiento natural. Una variable con estas propiedades se denomina variable ordinal, mientras que una variable categórica regular sin este tipo de ordenamiento especial se denomina variable nominal. Para simplificar los análisis, cualquier variable ordinal en este libro se tratará como una variable categórica nominal (no ordenada).

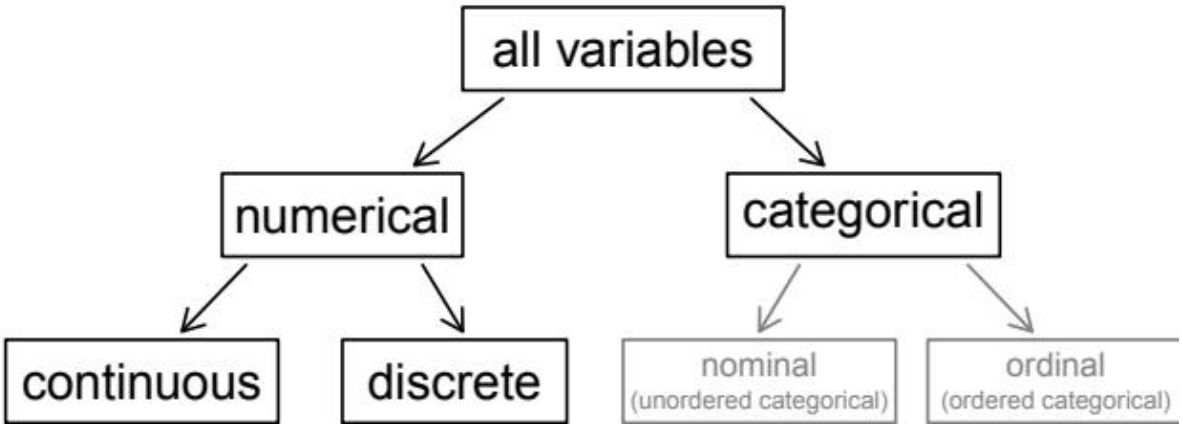


Figura 1.7: Desglose de las variables en sus respectivos tipos.

EJEMPLO 1.5

Se recolectaron datos sobre estudiantes en un curso de estadística. Se registraron tres variables para cada estudiante: número de hermanos, altura del estudiante y si el estudiante había

tomado previamente un curso de estadística. Clasifique cada una de las variables como numérica continua, numérica discreta o categórica.

El número de hermanos y la altura del estudiante representan variables numéricas. Debido a que el número de hermanos es un conteo, es discreto. La altura varía continuamente, por lo que es una variable numérica continua. La última variable clasifica a los estudiantes en dos categorías: aquellos que han tomado y aquellos que no han tomado un curso de estadística, lo que convierte a esta variable en categórica.

PRÁCTICA GUIADA 1.6

Un experimento está evaluando la efectividad de un nuevo fármaco en el tratamiento de las migrañas. Se utiliza una variable de grupo para indicar el grupo experimental para cada paciente: tratamiento o control. La variable num_migrañas representa el número de migrañas que el paciente experimentó durante un período de 3 meses. Clasifique cada variable como numérica o categórica.7

7La variable de grupo solo puede tomar uno de dos nombres de grupo, lo que la convierte en categórica. La variable num_migrañas describe un recuento del número de migrañas, que es un resultado donde la aritmética básica es sensata, lo que significa que este es un resultado numérico; más específicamente, dado que representa un recuento, num_migrañas es una variable numérica discreta.

1.2.3 Relaciones entre variables

Muchos análisis están motivados por un investigador que busca una relación entre dos o más variables. A un científico social le gustaría responder algunas de las siguientes preguntas:

- (1) Si la propiedad de vivienda es inferior al promedio nacional en un condado, ¿el porcentaje de estructuras de unidades múltiples en ese condado tenderá a estar por encima o por debajo del promedio nacional?
- (2) ¿Un aumento superior al promedio en la población del condado tiende a corresponder a condados con ingresos medios del hogar más altos o más bajos?
- (3) ¿Qué tan útil es el nivel de educación medio como predictor del ingreso medio del hogar para los condados de EE. UU.?

Para responder a estas preguntas, se deben recopilar datos, como el conjunto de datos del condado que se muestra en la Figura 1.5. El examen de las estadísticas resumidas podría proporcionar información sobre cada una de las tres preguntas sobre los condados. Además, se pueden utilizar gráficos para explorar visualmente los datos.

Los diagramas de dispersión son un tipo de gráfico que se utiliza para estudiar la relación entre dos variables numéricas. La Figura 1.8 compara las variables propiedad de vivienda y multi_unidad, que es el porcentaje de unidades en estructuras de unidades múltiples (por ejemplo, apartamentos, condominios). Cada punto en el gráfico representa un solo condado. Por ejemplo, el punto resaltado corresponde al condado 413 en el conjunto de datos del condado: el condado de Chattahoochee, Georgia, que tiene el 39,4% de las unidades en estructuras de unidades múltiples y una tasa de propiedad de vivienda del 31,3%. El diagrama de dispersión sugiere una relación entre las dos variables: los condados con una tasa más alta de unidades múltiples tienden a tener tasas de propiedad de vivienda más bajas. Podríamos hacer una lluvia de ideas sobre por qué existe esta relación e investigar cada idea para determinar cuáles son las explicaciones más razonables.

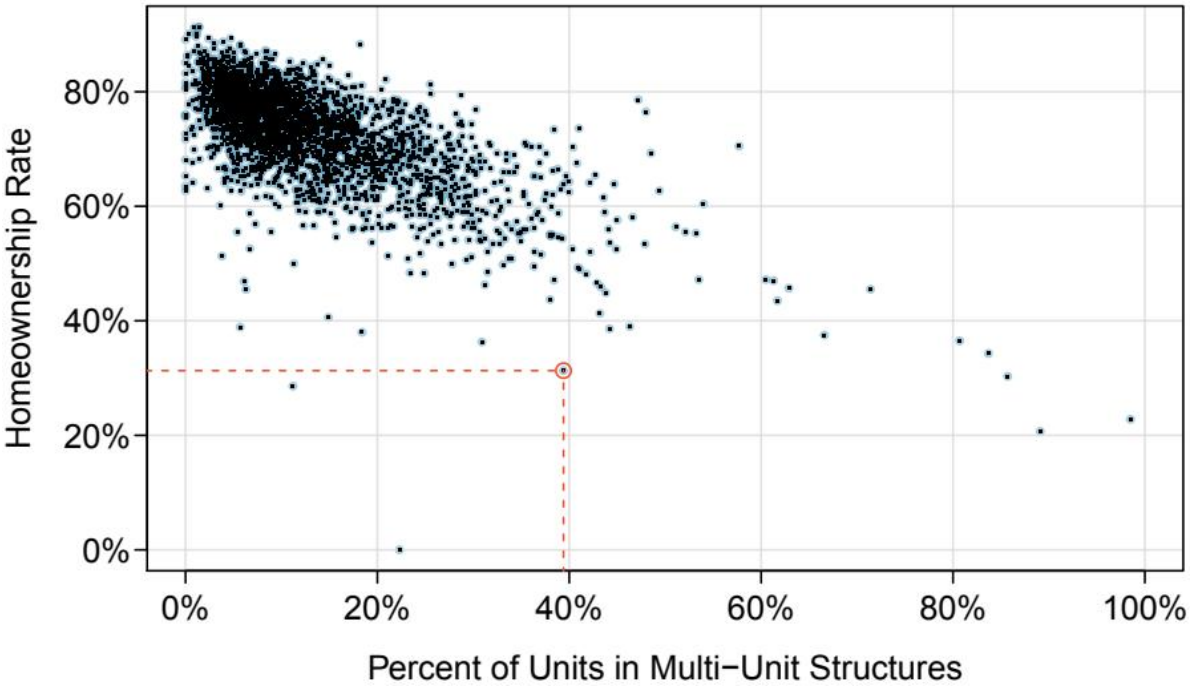


Figura 1.8: Un diagrama de dispersión de la propiedad de vivienda frente al porcentaje de unidades que se encuentran en estructuras de unidades múltiples para los condados de EE. UU. El punto resaltado representa el condado de Chattahoochee, Georgia, que tiene una tasa de unidades múltiples del 39,4% y una tasa de propiedad de vivienda del 31,3%.

Se dice que las tasas de multi_unidad y propiedad de vivienda están asociadas porque el gráfico muestra un patrón discernible. Cuando dos variables muestran alguna conexión entre sí, se denominan variables asociadas. Las variables asociadas también pueden denominarse variables dependientes y viceversa.

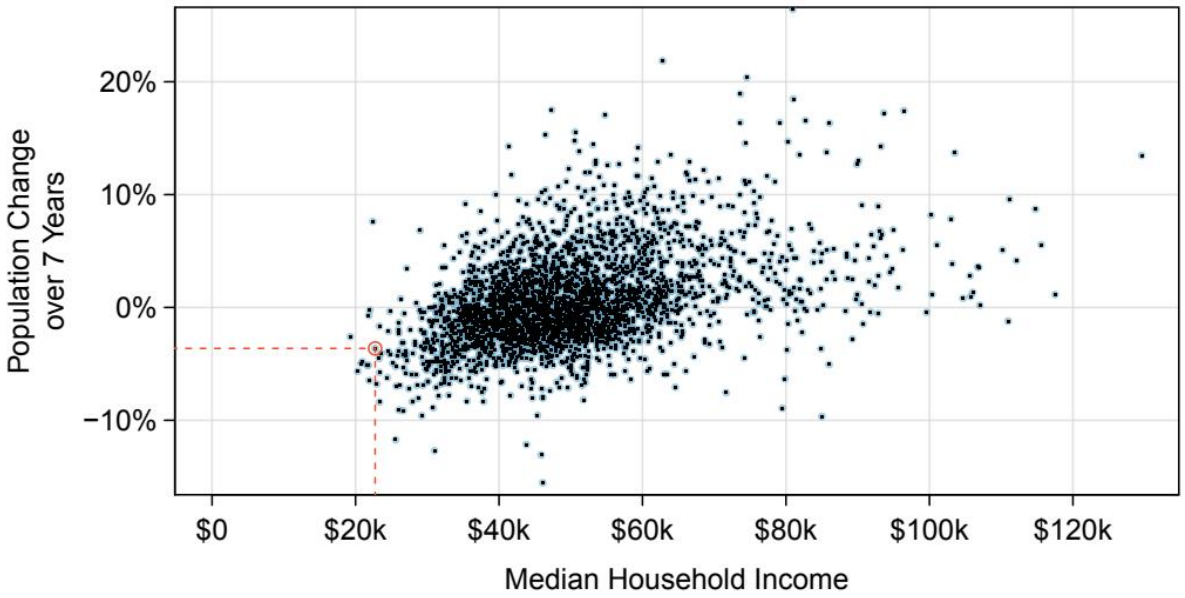


Figura 1.9: Un diagrama de dispersión que muestra el cambio de población frente al ingreso medio del hogar. Se destaca el condado de Owsley de Kentucky, que perdió el 3,63% de su población entre 2010 y 2017 y tenía un ingreso medio del hogar de $22,736.

Examine las variables en el conjunto de datos loan50, que se describen en la Figura 1.4 en la página 12. Cree dos preguntas sobre posibles relaciones entre variables en loan50 que sean de su interés.8

EJEMPLO 1.8

Este ejemplo examina la relación entre el cambio de población de un condado desde 2010 a 2017 y el ingreso medio por hogar, que se visualiza como un diagrama de dispersión en la Figura 1.9. ¿Están asociadas estas variables?

Cuanto mayor es el ingreso medio por hogar para un condado, mayor es el crecimiento de la población observado para el condado. Si bien esta tendencia no es cierta para todos los condados, la tendencia en el gráfico es evidente. Dado que existe alguna relación entre las variables, están asociadas.

Debido a que hay una tendencia a la baja en la Figura 1.8 – los condados con más unidades en estructuras de unidades múltiples están asociados con una menor propiedad de vivienda – se dice que estas variables están negativamente asociadas. Se muestra una asociación positiva en la relación entre el ingreso medio por hogar y el cambio de población en la Figura 1.9, donde los condados con un ingreso medio por hogar más alto tienden a tener tasas más altas de crecimiento de la población.

Si dos variables no están asociadas, entonces se dice que son independientes. Es decir, dos variables son independientes si no hay una relación evidente entre las dos.

ASOCIADAS O INDEPENDIENTES, NO AMBAS

Un par de variables están relacionadas de alguna manera (asociadas) o no (independientes). Ningún par de variables es a la vez asociado e independiente.

8Dos preguntas de ejemplo: (1) ¿Cuál es la relación entre el monto del préstamo y el ingreso total? (2) Si el ingreso de alguien está por encima del promedio, ¿su tasa de interés tenderá a estar por encima o por debajo del promedio?

1.2.4 Variables explicativas y de respuesta

Cuando hacemos preguntas sobre la relación entre dos variables, a veces también queremos determinar si el cambio en una variable causa un cambio en la otra. Considere la siguiente reformulación de una pregunta anterior sobre el conjunto de datos del condado:

Si hay un aumento en el ingreso medio por hogar en un condado, ¿esto impulsa un aumento en su población?

En esta pregunta, estamos preguntando si una variable afecta a otra. Si esta es nuestra creencia subyacente, entonces el ingreso medio por hogar es la variable explicativa y el cambio de población es la variable de respuesta en la relación hipotética.9

VARIABLES EXPLICATIVAS Y DE RESPUESTA

Cuando sospechamos que una variable podría afectar causalmente a otra, etiquetamos la primera variable como la variable explicativa y la segunda variable como la variable de respuesta.

> la variable explicativa podría afectar a la variable de respuesta

Para muchos pares de variables, no existe una relación hipotética, y estas etiquetas no se aplicarían a ninguna de las variables en tales casos.

Tenga en cuenta que el acto de etiquetar las variables de esta manera no garantiza que exista una relación causal. Una evaluación formal para verificar si una variable causa un cambio en otra requiere un experimento.

1.2.5 Introducción a los estudios observacionales y experimentos

Existen dos tipos principales de recolección de datos: estudios observacionales y experimentos.

Los investigadores realizan un estudio observacional cuando recolectan datos de una manera que no interfiere directamente con la forma en que surgen los datos. Por ejemplo, los investigadores pueden recolectar información a través de encuestas, revisar registros médicos o de empresas, o seguir a una cohorte de muchos individuos similares para formular hipótesis sobre por qué podrían desarrollarse ciertas enfermedades. En cada una de estas situaciones, los investigadores simplemente observan los datos que surgen. En general, los estudios observacionales pueden proporcionar evidencia de una asociación natural entre variables, pero no pueden por sí solos mostrar una conexión causal.

Cuando los investigadores quieren investigar la posibilidad de una conexión causal, realizan un experimento. Por lo general, habrá una variable explicativa y una variable de respuesta. Por ejemplo, podemos sospechar que la administración de un fármaco reducirá la mortalidad en pacientes con ataque cardíaco durante el año siguiente. Para verificar si realmente existe una conexión causal entre la variable explicativa y la respuesta, los investigadores recolectarán una muestra de individuos y los dividirán en grupos. A los individuos de cada grupo se les asigna un tratamiento. Cuando los individuos son asignados aleatoriamente a un grupo, el experimento se denomina experimento aleatorizado. Por ejemplo, cada paciente con ataque cardíaco en el ensayo del fármaco podría asignarse aleatoriamente, tal vez lanzando una moneda al aire, en uno de dos grupos: el primer grupo recibe un placebo (tratamiento falso) y el segundo grupo recibe el fármaco. Consulte el estudio de caso en la Sección 1.1 para ver otro ejemplo de un experimento, aunque ese estudio no empleó un placebo.

ASOCIACIÓN = CAUSACIÓN

En general, la asociación no implica causalidad, y la causalidad solo puede inferirse de un experimento aleatorizado.

9A veces, la variable explicativa se denomina variable independiente y la variable de respuesta se denomina variable dependiente. Sin embargo, esto se vuelve confuso ya que un par de variables podrían ser independientes o dependientes, por lo que evitamos este lenguaje.

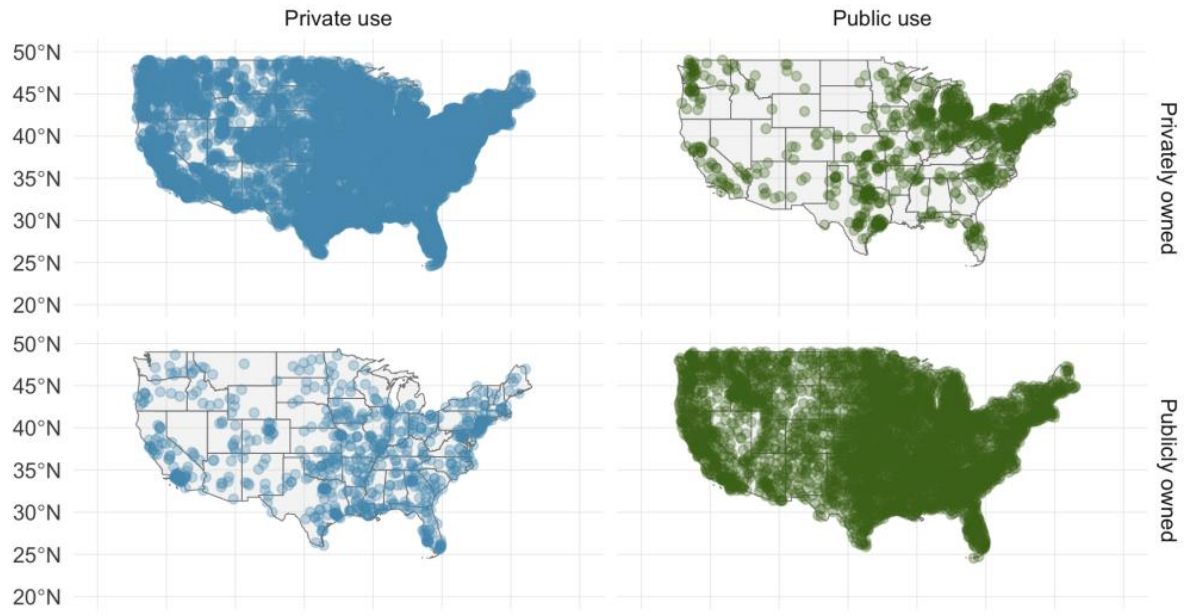
Ejercicios 1.3 Contaminación del aire y resultados del parto, componentes del estudio. Los investigadores recopilaron datos para examinar la relación entre los contaminantes del aire y los partos prematuros en el sur de California. Durante el estudio, los niveles de contaminación del aire se midieron mediante estaciones de monitoreo de la calidad del aire. Específicamente, se registraron los niveles de monóxido de carbono en partes por millón, dióxido de nitrógeno y ozono en partes por cada cien millones, y materia particulada gruesa (PM10) en µg/m3. Los datos de la duración de la gestación se recopilaron en 143.196 partos entre los años 1989 y 1993, y la exposición a la contaminación del aire durante la gestación se calculó para cada parto. El análisis sugirió que el aumento de PM10

y, en menor medida, las concentraciones de CO pueden estar asociados con la aparición de partos prematuros.10 - (a) Identifique la pregunta de investigación principal del estudio. - (b) ¿Quiénes son los sujetos en este estudio, y cuántos están incluidos? - (c) ¿Cuáles son las variables en el estudio? Identifique cada variable como numérica o categórica. Si numérica, indique si la variable es discreta o continua. Si categórica, indique si la variable es ordinal. **1.4 Método Buteyko, componentes del estudio.** El método Buteyko es una técnica de respiración superficial desarrollada por Konstantin Buteyko, un médico ruso, en 1952. La evidencia anecdótica sugiere que el método Buteyko puede reducir los síntomas del asma y mejorar la calidad de vida. En un estudio científico para determinar la efectividad de este método, los investigadores reclutaron a 600 pacientes con asma de entre 18 y 69 años que dependían de la medicación para el tratamiento del asma. Estos pacientes fueron divididos aleatoriamente en dos grupos de investigación: uno practicó el método Buteyko y el otro no. Los pacientes fueron evaluados en calidad de vida, actividad, síntomas de asma y reducción de medicación en una escala del 0 al 10. En promedio, los participantes del grupo Buteyko experimentaron una reducción significativa en los síntomas del asma y una mejora en la calidad de vida.11 - (a) Identifique la pregunta de investigación principal del estudio. - (b) ¿Quiénes son los sujetos en este estudio, y cuántos están incluidos? - (c) ¿Cuáles son las variables en el estudio? Identifique cada variable como numérica o categórica. Si numérica, indique si la variable es discreta o continua. Si categórica, indique si la variable es ordinal. **1.5 Tramposos, componentes del estudio.** Los investigadores que estudiaban la relación entre la honestidad, la edad y el autocontrol llevaron a cabo un experimento con 160 niños de entre 5 y 15 años. Los participantes informaron su edad, sexo y si eran hijo único o no. Los investigadores le pidieron a cada niño que lanzara una moneda justa en privado y que registrara el resultado (blanco o negro) en una hoja de papel, y dijeron que solo recompensarían a los niños que informaran blanco. Los hallazgos del estudio se pueden resumir de la siguiente manera: "La mitad de los estudiantes recibieron explícitamente la instrucción de no hacer trampa y los demás no recibieron ninguna instrucción explícita. En el grupo sin instrucciones, la probabilidad de hacer trampa fue uniforme en los grupos según las características del niño. En el grupo que recibió explícitamente la instrucción de no hacer trampa, las niñas fueron menos propensas a hacer trampa, y mientras que la tasa de hacer trampa no varió según la edad para los niños, disminuyó con la edad para las niñas."12 - (a) Identifique la pregunta de investigación principal del estudio. - (b) ¿Quiénes son los sujetos en este estudio, y cuántos están incluidos? - (c) ¿Cuántas variables se registraron para cada sujeto en el estudio para llegar a estas conclusiones? Indique las variables y sus tipos. **1.6 Ladrones, componentes del estudio.** En un estudio sobre la relación entre la clase socioeconómica y el comportamiento poco ético, 129 estudiantes universitarios de la Universidad de California en Berkeley fueron invitados a identificarse como de clase social baja o alta al compararse con otros con el dinero más (menos), la educación más (menos), y los trabajos más (menos) respetados. También se les presentó un frasco de caramelos envueltos individualmente e informaron que los caramelos eran para niños en un laboratorio cercano, pero que podían tomar algunos si querían. Después de completar algunas tareas sin relación, los participantes informaron la cantidad de caramelos que habían tomado.13 - (a) Identifique la pregunta de investigación principal del estudio. - (b) ¿Quiénes son los sujetos en este estudio, y cuántos

están incluidos? - (c) El estudio encontró que los estudiantes que se identificaron como de clase alta tomaron más caramelos que otros. ¿Cuántas variables se registraron para cada sujeto en el estudio para llegar a estas conclusiones? Indique las variables y sus tipos. **1.7 Migraña y acupuntura, Parte II.** El ejercicio 1.1 introdujo un estudio que exploraba si la acupuntura tenía algún efecto en las migrañas. Los investigadores llevaron a cabo un estudio controlado aleatorio en el que los pacientes fueron asignados aleatoriamente a uno de dos grupos: tratamiento o control. Los pacientes del grupo de tratamiento recibieron acupuntura diseñada específicamente para tratar las migrañas. Los pacientes del grupo de control recibieron acupuntura placebo (inserción de agujas en lugares no acupunturales). 24 horas después de que los pacientes recibieron acupuntura, se les preguntó si estaban libres de dolor. ¿Cuáles son las variables explicativas y de respuesta en este estudio? **1.8 Sinusitis y antibióticos, Parte II.** El ejercicio [1.2](#page-10-2] introdujo un estudio que exploraba si los antibióticos tenían algún efecto en la sinusitis. Los investigadores reclutaron a pacientes con sinusitis y los asignaron aleatoriamente a un grupo que recibió antibióticos y a un grupo que recibió un placebo. Después de una semana, los investigadores evaluaron a los pacientes para ver si sus síntomas habían mejorado. ¿Cuáles son las variables explicativas y de respuesta en este estudio? **1.9 El efecto del ejercicio en el estado de ánimo, componentes del estudio.** Un investigador está interesado en determinar si el ejercicio afecta el estado de ánimo. El investigador recluta a 30 voluntarios y les pide que informen su estado de ánimo en una escala del 0 al 10, donde 0 es el estado de ánimo más bajo y 10 es el estado de ánimo más alto. Luego, el investigador pide a los participantes que hagan ejercicio durante 30 minutos y luego informa su estado de ánimo nuevamente. ¿Cuál es la variable dependiente en este estudio? **1.10 El efecto de la música en el rendimiento académico, componentes del estudio.** Un profesor está interesado en determinar si la música afecta el rendimiento académico. El profesor recluta a 50 estudiantes y les pide que tomen un examen. Luego, el profesor pide a los estudiantes que escuchen música durante 30 minutos y luego les pide que tomen otro examen. ¿Cuál es la variable independiente en este estudio? 15Centro Nacional de STEM, Grandes conjuntos de datos de stats4schools.

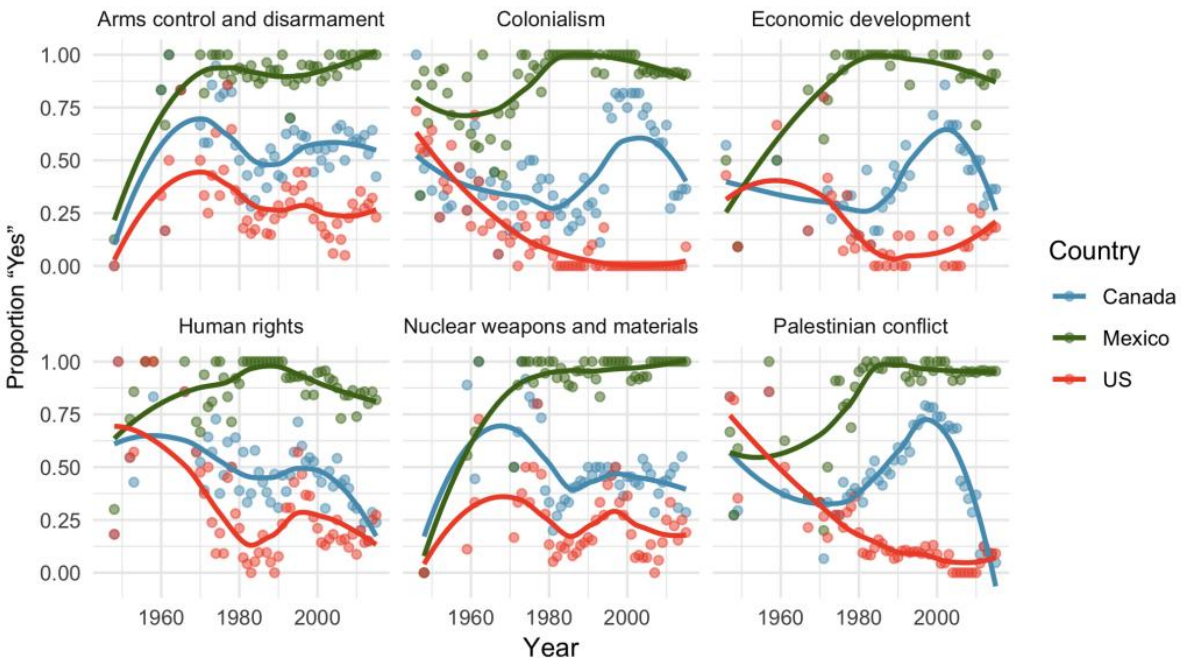
1.2. BASES DE DATOS 21

1.11 Aeropuertos de EE.UU. La visualización a continuación muestra la distribución geográfica de los aeropuertos en los Estados Unidos contiguos y Washington, DC. Esta visualización se construyó a partir de un conjunto de datos donde cada observación es un aeropuerto.16



-
- (a) Enumere las variables utilizadas para crear esta visualización.
- (b) Indique si cada variable en el estudio es numérica o categórica. Si es numérica, identifique si es continua o discreta. Si es categórica, indique si la variable es ordinal.

1.12 Votos de la ONU. La visualización a continuación muestra los patrones de votación en los Estados Unidos, Canadá y México en la Asamblea General de las Naciones Unidas sobre una variedad de temas. Específicamente, para un año dado entre 1946 y 2015, muestra el porcentaje de votaciones nominales en las que el país votó sí para cada tema. Esta visualización se construyó a partir de un conjunto de datos donde cada observación es un par país/año.17



- (a) Enumere las variables utilizadas para crear esta visualización.
- (b) Indique si cada variable en el estudio es numérica o categórica. Si es numérica, identifique si es continua o discreta. Si es categórica, indique si la variable es ordinal.

16Administración Federal de Aviación, www.faa.gov/airports/airport safety/airportdata 5010.

17David Robinson. unvotes: Datos de votación de la Asamblea General de las Naciones Unidas. Paquete R versión 0.2.0. 2017. url: https://CRAN.R-project.org/package=unvotes.

1.3 Principios y estrategias de muestreo

El primer paso para llevar a cabo una investigación es identificar los temas o preguntas que se investigarán. Una pregunta de investigación claramente definida es útil para identificar qué sujetos o casos deben estudiarse y qué variables son importantes. También es importante considerar cómo se recopilan los datos para que sean confiables y ayuden a lograr los objetivos de la investigación.

1.3.1 Poblaciones y muestras

Considera las siguientes tres preguntas de investigación:

- 1. ¿Cuál es el contenido promedio de mercurio en el pez espada en el Océano Atlántico?
- 2. Durante los últimos 5 años, ¿cuál es el tiempo promedio para completar una licenciatura para los estudiantes de Duke?
- 3. ¿Un nuevo fármaco reduce el número de muertes en pacientes con enfermedad cardíaca grave?

Cada pregunta de investigación se refiere a una población objetivo. En la primera pregunta, la población objetivo son todos los peces espada en el Océano Atlántico, y cada pez representa un caso. A menudo, es demasiado costoso recopilar datos para cada caso en una población. En cambio, se toma una muestra. Una muestra representa un subconjunto de los casos y a menudo es una pequeña fracción de la población. Por ejemplo, se podrían seleccionar 60 peces espada (o algún otro número) de la población, y estos datos de muestra podrían usarse para proporcionar una estimación del promedio de la población y responder a la pregunta de investigación.

PRÁCTICA GUIADA 1.9

Para la segunda y tercera pregunta anteriores, identifica la población objetivo y qué representa un caso individual.18

1.3.2 Evidencia anecdótica

Considera las siguientes posibles respuestas a las tres preguntas de investigación:

- 1. Un hombre en las noticias se intoxicó con mercurio por comer pez espada, por lo que la concentración promedio de mercurio en el pez espada debe ser peligrosamente alta.
- 2. Conocí a dos estudiantes que tardaron más de 7 años en graduarse de Duke, por lo que debe tomar más tiempo graduarse en Duke que en muchas otras universidades.
- 3. El papá de mi amigo tuvo un ataque al corazón y murió después de que le dieron un nuevo medicamento para la enfermedad cardíaca, por lo que el medicamento no debe funcionar.

Cada conclusión se basa en datos. Sin embargo, hay dos problemas. Primero, los datos solo representan uno o dos casos. Segundo, y más importante, no está claro si estos casos son realmente representativos de la población. Los datos recopilados de esta manera descuidada se denominan evidencia anecdótica.

EVIDENCIA ANECDÓTICA

Ten cuidado con los datos recopilados de forma descuidada. Tal evidencia puede ser verdadera y verificable, pero solo puede representar casos extraordinarios.

18(2) La primera pregunta solo es relevante para los estudiantes que completan su título; el promedio no se puede calcular utilizando a un estudiante que nunca terminó su título. Por lo tanto, solo los estudiantes de pregrado de Duke que se graduaron en los últimos cinco años representan casos en la población en consideración. Cada uno de estos estudiantes es un caso individual. (3) Una persona con una enfermedad cardíaca grave representa un caso. La población incluye a todas las personas con una enfermedad cardíaca grave.

Figura 1.10: En febrero de 2010, algunos expertos de los medios citaron una gran tormenta de nieve como evidencia válida contra el calentamiento global. Como señaló el comediante Jon Stewart, "Es una tormenta, en una región, de un país".

La evidencia anecdótica típicamente se compone de casos inusuales que recordamos en función de sus características sorprendentes. Por ejemplo, es más probable que recordemos a las dos personas que conocimos que tardaron 7 años en graduarse que a las otras seis que se graduaron en cuatro años. En lugar de observar los casos más inusuales, deberíamos examinar una muestra de muchos casos que representen a la población.

1.3.3 Muestreo de una población

Podríamos intentar estimar el tiempo hasta la graduación de los estudiantes de pregrado de Duke en los últimos 5 años recolectando una muestra de estudiantes. Todos los graduados en los últimos 5 años representan la población, y los graduados que son seleccionados para su revisión se denominan colectivamente la muestra. En general, siempre buscamos seleccionar aleatoriamente una muestra de una población. El tipo más básico de selección aleatoria es equivalente a cómo se realizan las rifas. Por ejemplo, al seleccionar graduados, podríamos escribir el nombre de cada graduado en un boleto de rifa y sacar 100 boletos. Los nombres seleccionados representarían una muestra aleatoria de 100 graduados. Elegimos muestras aleatoriamente para reducir la posibilidad de introducir sesgos.

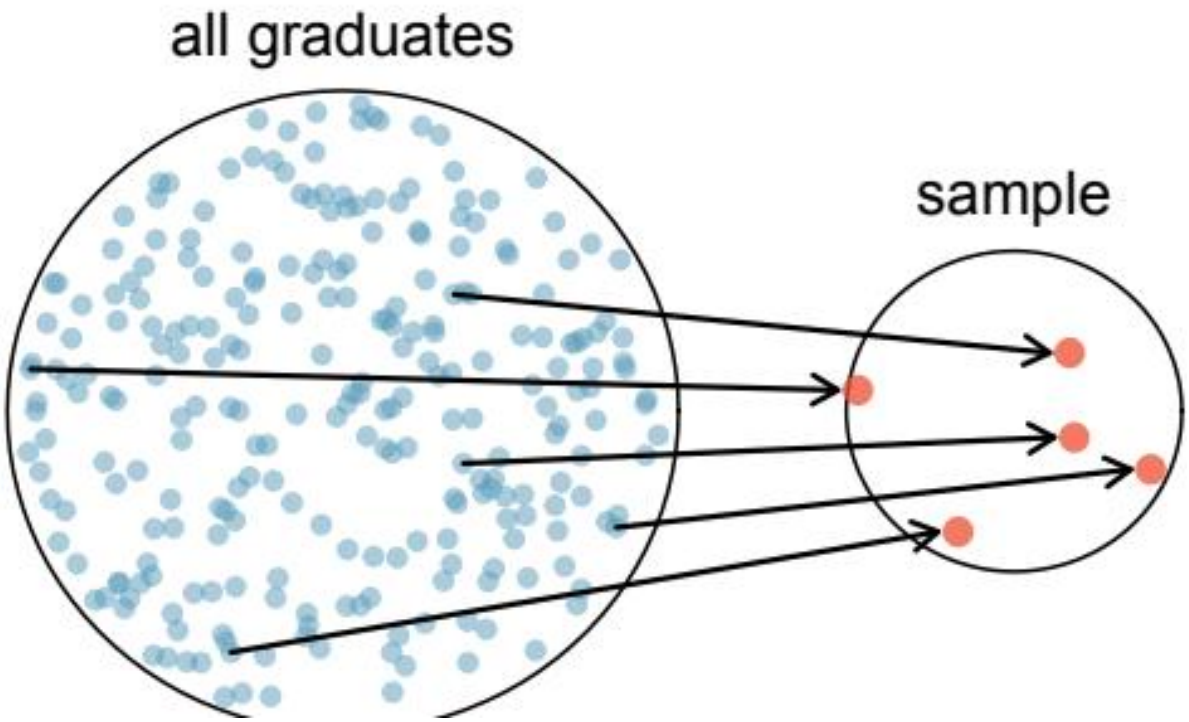


Figura 1.11: En este gráfico, cinco graduados son seleccionados aleatoriamente de la población para ser incluidos en la muestra.

EJEMPLO 1.10

Supongamos que le pedimos a un estudiante que se especializa en nutrición que seleccione a varios graduados para el estudio. ¿Qué tipo de estudiantes crees que podría reunir? ¿Crees que su muestra sería representativa de todos los graduados?

Tal vez elegiría un número desproporcionado de graduados de campos relacionados con la salud. O tal vez su selección sería una buena representación de la población. Al seleccionar muestras a mano, corremos el riesgo de elegir una muestra sesgada, incluso si su sesgo no es intencionado.

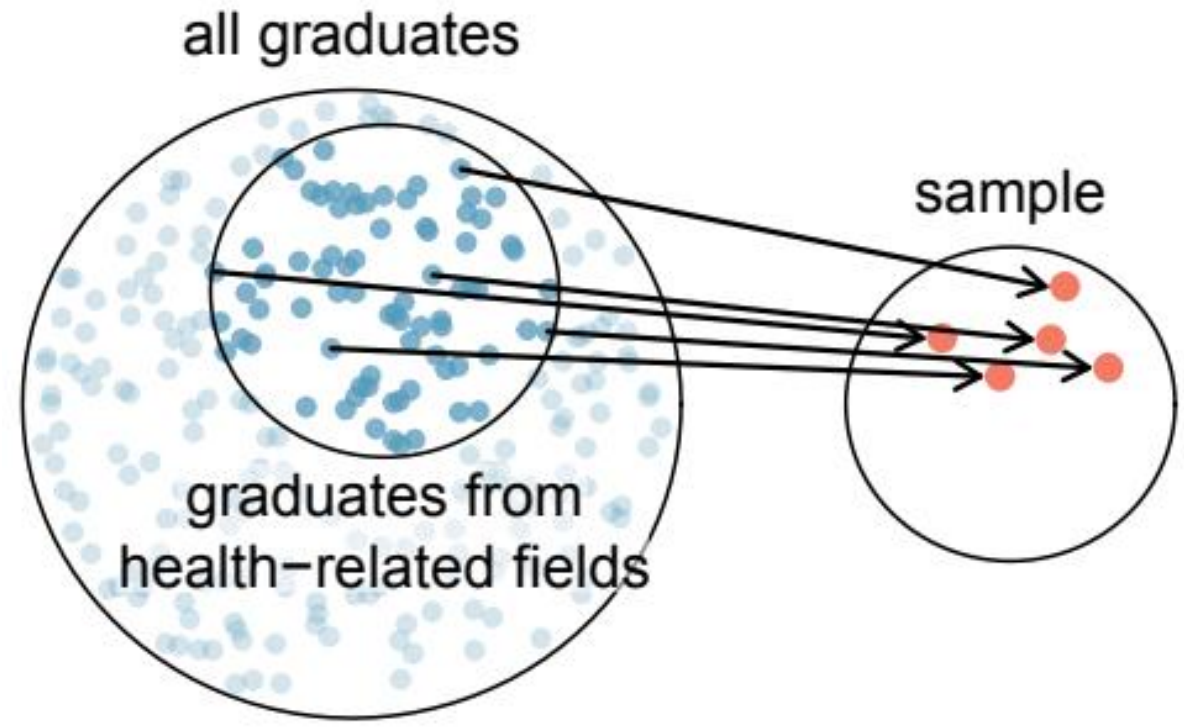


Figura 1.12: Si se le pide que elija una muestra de graduados, un estudiante de nutrición podría elegir inadvertidamente un número desproporcionado de graduados de carreras relacionadas con la salud.

Si a alguien se le permitiera elegir exactamente qué graduados se incluyen en la muestra, es muy posible que la muestra se inclinara hacia los intereses de esa persona, lo que puede ser completamente involuntario. Esto introduce sesgo en una muestra. El muestreo aleatorio ayuda a resolver este problema. La muestra aleatoria más básica se llama muestra aleatoria simple,

que es equivalente a usar una rifa para seleccionar casos. Esto significa que cada caso en la población tiene la misma probabilidad de ser incluido y no hay una conexión implícita entre los casos en la muestra.

El acto de tomar una muestra aleatoria simple ayuda a minimizar el sesgo. Sin embargo, el sesgo puede surgir de otras maneras. Incluso cuando las personas son elegidas al azar, por ejemplo, para encuestas, se debe tener precaución si la tasa de no respuesta es alta. Por ejemplo, si solo el 30% de las personas seleccionadas al azar para una encuesta responden, entonces no está claro si los resultados son representativos de toda la población. Este sesgo de no respuesta puede sesgar los resultados.

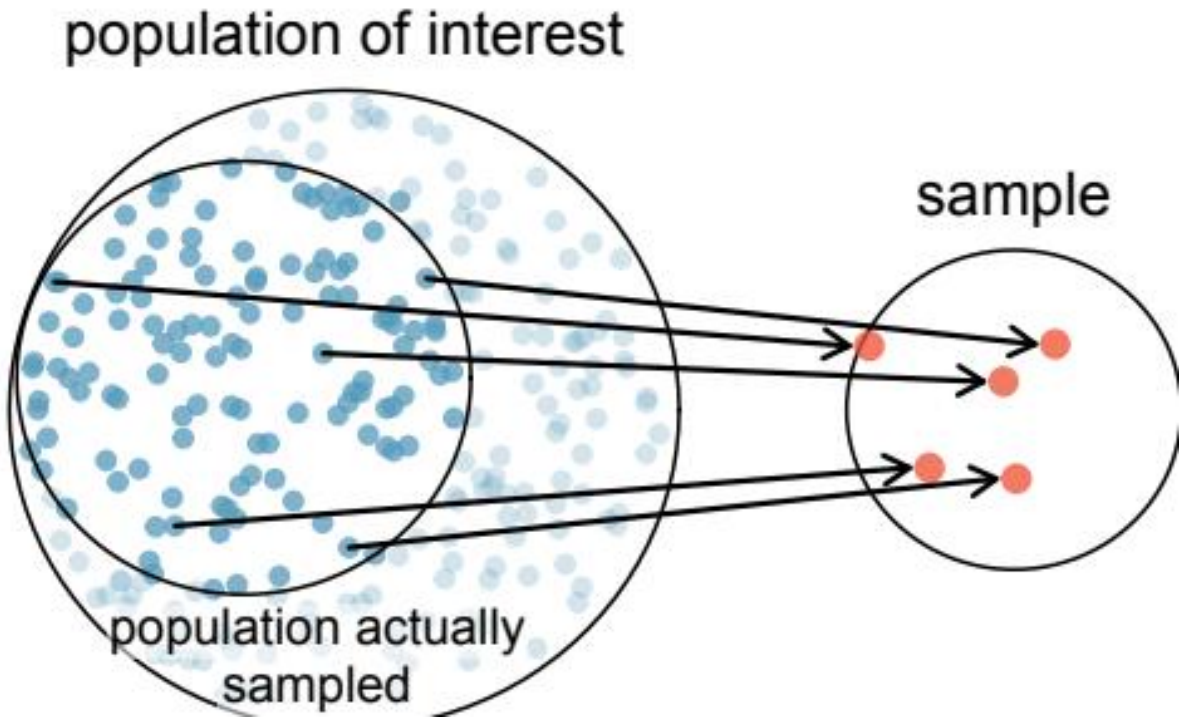


Figura 1.13: Debido a la posibilidad de no respuesta, los estudios de encuestas pueden llegar solo a un cierto grupo dentro de la población. Es difícil, y muchas veces imposible, solucionar por completo este problema.

Otro error común es una muestra de conveniencia, donde los individuos que son fácilmente accesibles tienen más probabilidades de ser incluidos en la muestra. Por ejemplo, si una encuesta política se realiza deteniendo a personas que caminan en el Bronx, esto no representará a toda la ciudad de Nueva York. A menudo es difícil discernir qué subpoblación representa una muestra de conveniencia.

PRÁCTICA GUIADA 1.11

Podemos acceder fácilmente a las calificaciones de productos, vendedores y empresas a través de sitios web. Estas calificaciones se basan únicamente en aquellas personas que se esfuerzan por proporcionar una calificación. Si el 50% de las reseñas en línea de un producto son negativas, ¿cree que esto significa que el 50% de los compradores están insatisfechos con el producto?19

19Las respuestas variarán. A partir de nuestras propias experiencias anecdóticas, creemos que las personas tienden a despotricar más sobre los productos que no cumplieron con las expectativas que a entusiasmarse con aquellos que funcionan como se esperaba. Por esta razón, sospechamos que existe un sesgo negativo en las calificaciones de productos en sitios como Amazon. Sin embargo, dado que nuestras experiencias pueden no ser representativas, también mantenemos una mente abierta.

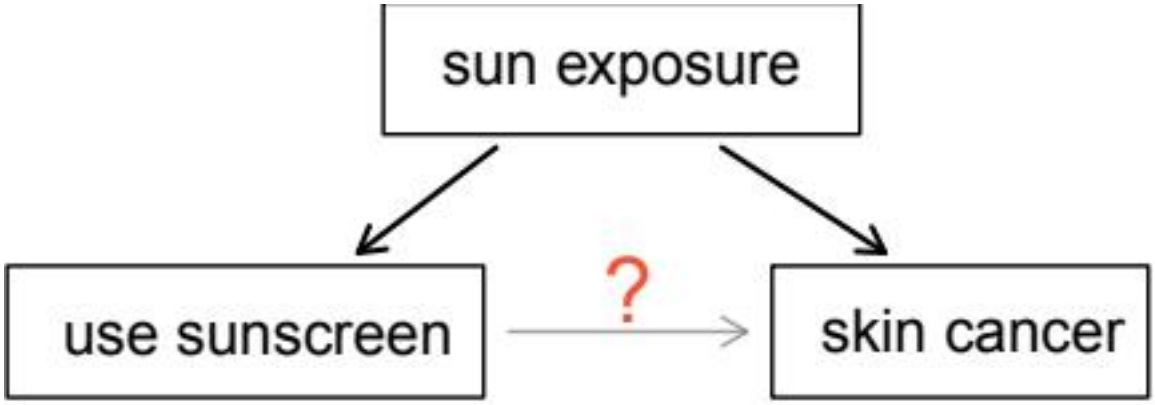
1.3.4 Estudios observacionales

Los datos donde no se ha aplicado explícitamente ningún tratamiento (o se ha retenido explícitamente) se denominan datos observacionales. Por ejemplo, los datos de préstamos y los datos del condado descritos en la Sección 1.2 son ejemplos de datos observacionales. Hacer conclusiones causales basadas en experimentos suele ser razonable. Sin embargo, hacer las mismas conclusiones causales basadas en datos observacionales puede ser traicionero y no se recomienda. Por lo tanto, los estudios observacionales generalmente solo son suficientes para mostrar asociaciones o formar hipótesis que luego verificamos utilizando experimentos.

PRÁCTICA GUIADA 1.12

Supongamos que un estudio observacional rastreó el uso de protector solar y el cáncer de piel, y se descubrió que cuanto más protector solar usaba alguien, más probable era que la persona tuviera cáncer de piel. ¿Significa esto que el protector solar causa cáncer de piel?20

Algunas investigaciones previas nos dicen que usar protector solar en realidad reduce el riesgo de cáncer de piel, por lo que tal vez haya otra variable que pueda explicar esta hipotética asociación entre el uso de protector solar y el cáncer de piel. Una pieza importante de información que está ausente es la exposición al sol. Si alguien está expuesto al sol todo el día, es más probable que use protector solar y que le dé cáncer de piel. La exposición al sol no se tiene en cuenta en la simple investigación.



La exposición al sol es lo que se llama una variable de confusión,21 que es una variable que está correlacionada tanto con las variables explicativas como con las variables de respuesta. Si bien un método para justificar la elaboración de conclusiones causales a partir de estudios observacionales es agotar la búsqueda de variables de confusión, no hay garantía de que todas las variables de confusión puedan examinarse o medirse.

PRÁCTICA GUIADA 1.13

La Figura 1.8 muestra una asociación negativa entre la tasa de propiedad de vivienda y el porcentaje de estructuras de unidades múltiples en un condado. Sin embargo, no es razonable concluir que existe una relación causal entre las dos variables. Sugiera una variable que pueda explicar la relación negativa.22

Los estudios observacionales se presentan en dos formas: estudios prospectivos y retrospectivos. Un estudio prospectivo identifica a los individuos y recopila información a medida que se desarrollan los eventos. Por ejemplo, los investigadores médicos pueden identificar y seguir a un grupo de pacientes durante muchos años para evaluar las posibles influencias del comportamiento en el riesgo de cáncer. Un ejemplo de tal estudio es el Estudio de Salud de las Enfermeras, que comenzó en 1976 y se amplió en 1989. Este estudio prospectivo recluta enfermeras registradas y luego recopila datos de ellas mediante cuestionarios. Los estudios retrospectivos recopilan datos después de que los eventos han tenido lugar, por ejemplo, los investigadores pueden revisar eventos pasados en registros médicos. Algunos conjuntos de datos pueden contener variables recopiladas tanto prospectiva como retrospectivamente.

1.3.5 Cuatro métodos de muestreo

Casi todos los métodos estadísticos se basan en la noción de aleatoriedad implícita. Si los datos observacionales no se recopilan de un marco aleatorio de una población, estos métodos estadísticos – las estimaciones y los errores asociados con las estimaciones – no son confiables. Aquí consideramos cuatro técnicas de muestreo aleatorio: muestreo simple, estratificado, por

conglomerados y multietápico. Las figuras 1.14 y 1.15 proporcionan representaciones gráficas de estas técnicas.

20No. Consulte el párrafo posterior al ejercicio para obtener una explicación.

21También llamada variable latente, factor de confusión o confusor.

22Las respuestas variarán. La densidad de población puede ser importante. Si un condado es muy denso, esto puede requerir que una mayor fracción de residentes viva en estructuras de varias unidades. Además, la alta densidad puede contribuir al aumento del valor de las propiedades, lo que hace que la propiedad de la vivienda sea inviable para muchos residentes.

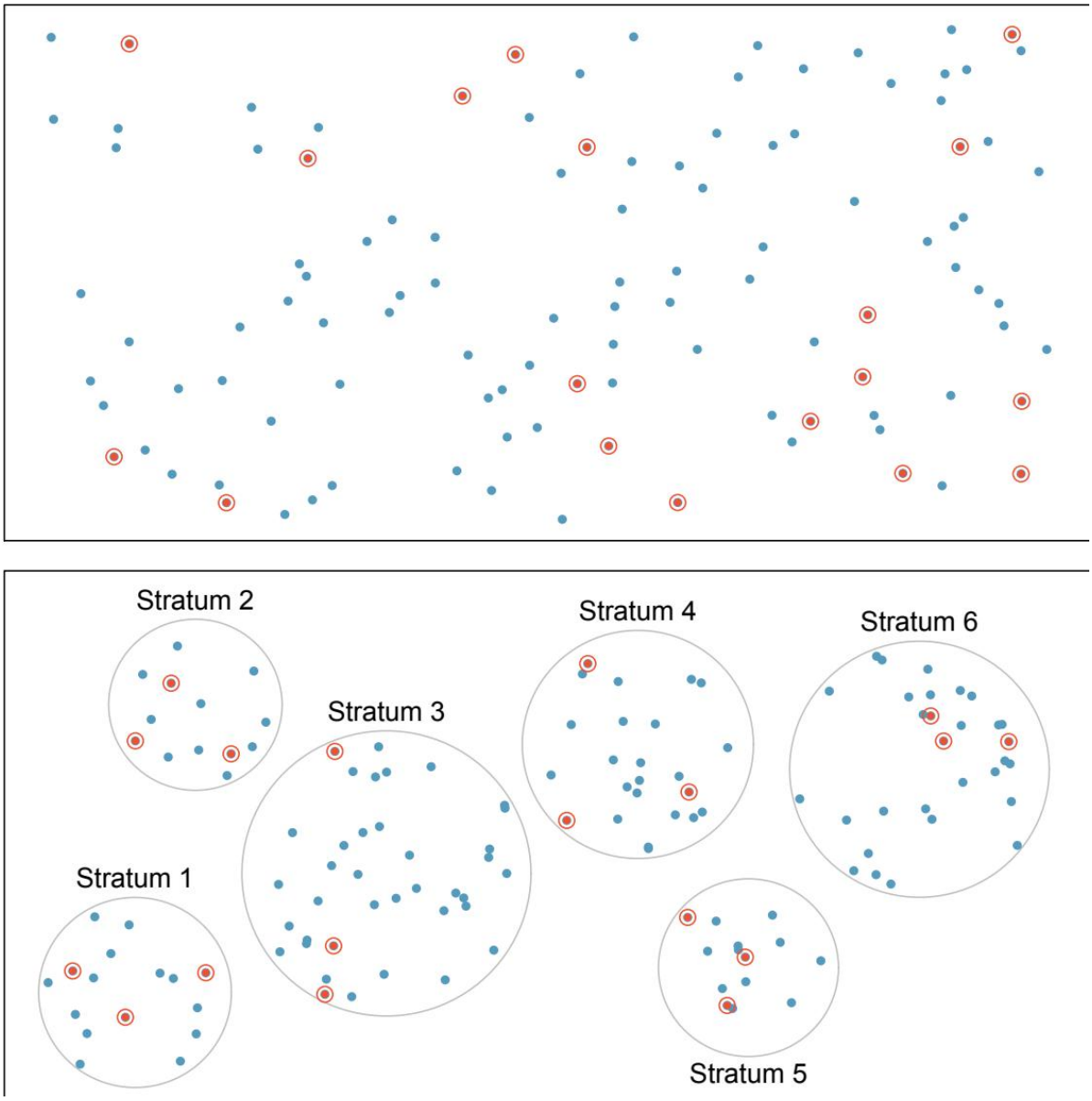


Figura 1.14: Ejemplos de muestreo aleatorio simple y estratificado. En el panel superior, se utilizó el muestreo aleatorio simple para seleccionar aleatoriamente los 18 casos. En el panel inferior, se utilizó el muestreo estratificado: los casos se agruparon en estratos, luego se empleó el muestreo aleatorio simple dentro de cada estrato.

El muestreo aleatorio simple es probablemente la forma más intuitiva de muestreo aleatorio. Considere los salarios de los jugadores de las Grandes Ligas de Béisbol (MLB), donde cada jugador es miembro de uno de los 30 equipos de la liga. Para tomar una muestra aleatoria simple de 120 jugadores de béisbol y sus salarios, podríamos escribir los nombres de los varios cientos de jugadores de esa temporada en trozos de papel, dejar caer los trozos en un balde, agitar el balde hasta que estemos seguros de que los nombres están mezclados, luego sacar los trozos hasta que tengamos la muestra de 120 jugadores. En general, una muestra se conoce como "aleatoria simple" si cada caso en la población tiene la misma probabilidad de ser incluido en la muestra final y saber que un caso está incluido en una muestra no proporciona información útil sobre qué otros casos están incluidos.

El muestreo estratificado es una estrategia de muestreo de divide y vencerás. La población se divide en grupos llamados estratos. Los estratos se eligen de modo que los casos similares se agrupen, luego se emplea un segundo método de muestreo, generalmente el muestreo aleatorio simple, dentro de cada estrato. En el ejemplo del salario del béisbol, los equipos podrían representar los estratos, ya que algunos equipos tienen mucho más dinero (¡hasta 4 veces más!). Entonces podríamos muestrear aleatoriamente a 4 jugadores de cada equipo para un total de 120 jugadores.

El muestreo estratificado es especialmente útil cuando los casos en cada estrato son muy similares con respecto al resultado de interés. La desventaja es que analizar datos de una muestra estratificada es una tarea más compleja que analizar datos de una muestra aleatoria simple. Los métodos de análisis introducidos en este libro deberían ampliarse para analizar los datos recopilados mediante el muestreo estratificado.

EJEMPLO 1.14

¿Por qué sería bueno que los casos dentro de cada estrato fueran muy similares?

Podríamos obtener una estimación más estable para la subpoblación en un estrato si los casos son muy similares, lo que conduciría a estimaciones más precisas dentro de cada grupo. Cuando combinamos estas estimaciones en una sola estimación para la población completa, esa estimación de la población tenderá a ser más precisa, ya que cada estimación de grupo individual es en sí misma más precisa.

En un muestreo por conglomerados, dividimos la población en muchos grupos, llamados conglomerados. Luego muestreamos un número fijo de conglomerados e incluimos todas las observaciones de cada uno de esos conglomerados en la muestra. Un muestreo multietápico es como un muestreo por conglomerados, pero en lugar de mantener todas las observaciones en cada conglomerado, recopilamos una muestra aleatoria dentro de cada conglomerado seleccionado.

A veces, el muestreo por conglomerados o multietápico puede ser más económico que las técnicas de muestreo alternativas. Además, a diferencia del muestreo estratificado, estos enfoques son más útiles cuando hay mucha variabilidad de caso a caso dentro de un conglomerado, pero los conglomerados en sí mismos no se ven muy diferentes entre sí. Por ejemplo, si los vecindarios representaran conglomerados, entonces el muestreo por conglomerados o multietápico funciona mejor cuando los vecindarios son muy diversos. Una desventaja de estos métodos es que normalmente se requieren técnicas más avanzadas para analizar los datos, aunque los métodos de este libro se pueden ampliar para manejar tales datos.

EJEMPLO 1.15

Supongamos que estamos interesados en estimar la tasa de malaria en una porción densamente tropical de la Indonesia rural. Nos enteramos de que hay 30 aldeas en esa parte de la jungla indonesia, cada una más o menos similar a la siguiente. Nuestro objetivo es examinar a 150 personas para detectar malaria. ¿Qué método de muestreo se debe emplear?

Un muestreo aleatorio simple probablemente extraerá individuos de las 30 aldeas, lo que podría hacer que la recopilación de datos sea extremadamente costosa. El muestreo estratificado sería un desafío ya que no está claro cómo construiríamos estratos de individuos similares. Sin embargo, el muestreo por conglomerados o el muestreo multietápico parecen ser muy buenas ideas. Si decidiéramos utilizar el muestreo multietápico, podríamos seleccionar aleatoriamente la mitad de las aldeas y luego seleccionar aleatoriamente a 10 personas de cada una. Esto probablemente reduciría sustancialmente nuestros costos de recopilación de datos en comparación con un muestreo aleatorio simple, y el muestreo por conglomerados aún nos brindaría información confiable, incluso si tuviéramos que analizar los datos con métodos un poco más avanzados de los que discutimos en este libro.

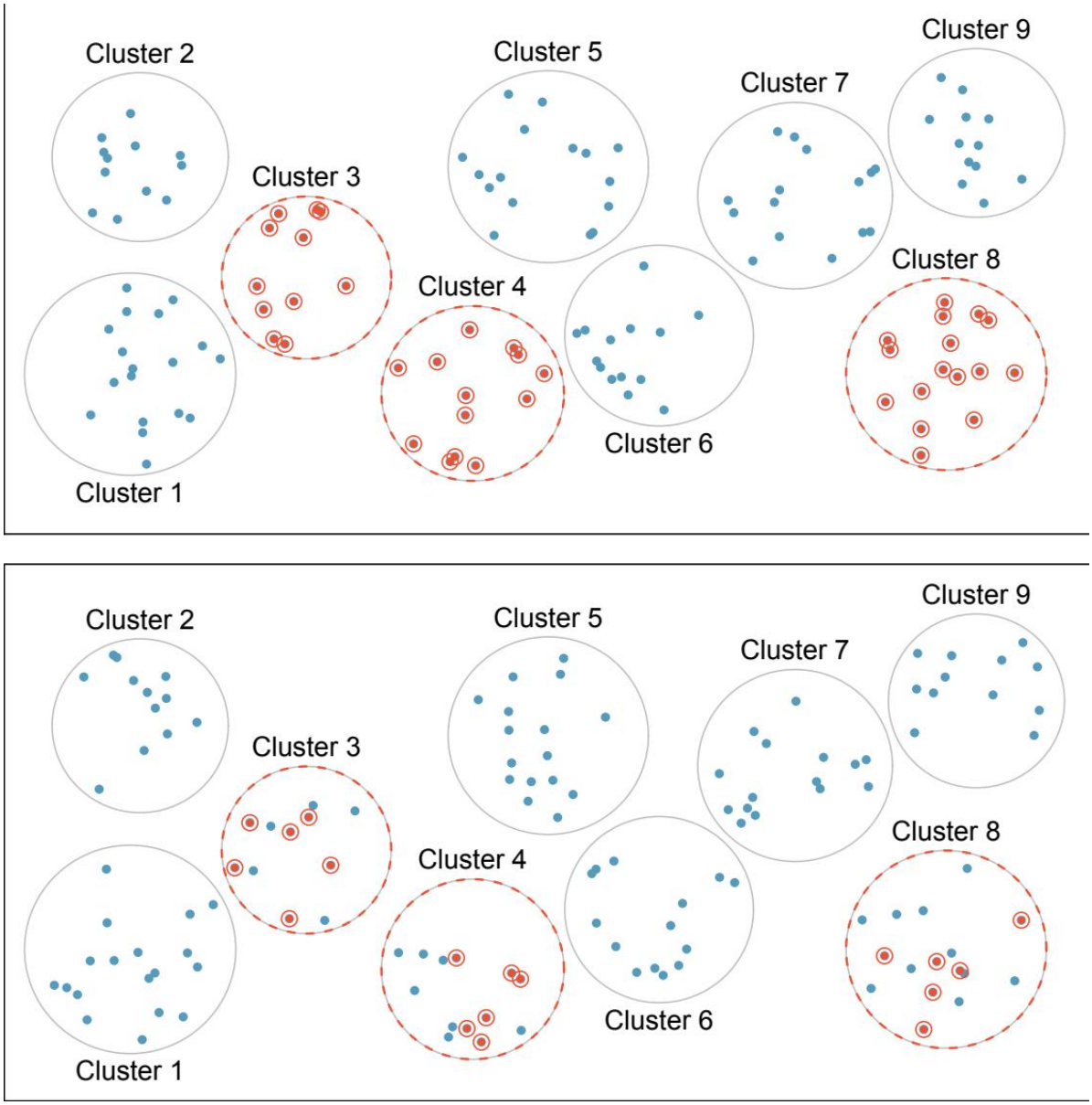


Figura 1.15: Ejemplos de muestreo por conglomerados y multietápico. En el panel superior, se utilizó el muestreo por conglomerados: los datos se agruparon en nueve conglomerados, se muestrearon tres de estos conglomerados y se incluyeron todas las observaciones dentro de estos tres conglomerados en la muestra. En el panel inferior, se utilizó el muestreo multietápico, que difiere del muestreo por conglomerados solo en que seleccionamos aleatoriamente un subconjunto de cada conglomerado para incluirlo en la muestra en lugar de medir cada caso en cada conglomerado muestreado.

1.4.1 Principios del diseño experimental

Los experimentos aleatorizados generalmente se basan en cuatro principios.

- Control. Los investigadores asignan tratamientos a los casos y hacen todo lo posible para controlar cualquier otra diferencia en los grupos.27 Por ejemplo, cuando los pacientes toman un medicamento en forma de píldora, algunos pacientes toman la píldora con solo un sorbo de agua, mientras que otros pueden tomarla con un vaso entero de agua. Para controlar el efecto del consumo de agua, un médico puede pedir a todos los pacientes que beban un vaso de 12 onzas de agua con la píldora.
- Aleatorización. Los investigadores aleatorizan a los pacientes en grupos de tratamiento para tener en cuenta las variables que no se pueden controlar. Por ejemplo, algunos pacientes pueden ser más susceptibles a una enfermedad que otros debido a sus hábitos alimenticios. La aleatorización de los pacientes en el grupo de tratamiento o control ayuda a igualar tales diferencias y también evita que sesgos accidentales entren en el estudio.
- Réplica. Cuantos más casos observen los investigadores, con mayor precisión podrán estimar el efecto de la variable explicativa sobre la respuesta. En un solo estudio, replicamos recogiendo una muestra suficientemente grande. Además, un grupo de científicos puede replicar un estudio completo para verificar un hallazgo anterior.
- Bloqueo. Los investigadores a veces saben o sospechan que variables, además del tratamiento, influyen en la respuesta. En estas circunstancias, primero pueden agrupar a los individuos en función de esta variable en bloques y luego aleatorizar los casos dentro de cada bloque a los grupos de tratamiento. Esta estrategia se conoce a menudo como bloqueo. Por ejemplo, si estamos observando el efecto de un fármaco en los ataques cardíacos, podríamos primero dividir a los pacientes en el estudio en bloques de bajo riesgo y alto riesgo, luego asignar aleatoriamente la mitad de los pacientes de cada bloque al grupo de control y la otra mitad al grupo de tratamiento, como se muestra en la Figura 1.16. Esta estrategia asegura que cada grupo de tratamiento tenga un número igual de pacientes de bajo riesgo y alto riesgo.

Es importante incorporar los tres primeros principios del diseño experimental en cualquier estudio, y este libro describe los métodos aplicables para analizar los datos de tales experimentos. El bloqueo es una técnica un poco más avanzada, y los métodos estadísticos de este libro pueden extenderse para analizar los datos recogidos mediante el bloqueo.

1.4.2 Reducción del sesgo en experimentos con humanos

Los experimentos aleatorizados son el estándar de oro para la recogida de datos, pero no garantizan una perspectiva imparcial sobre la relación causa y efecto en todos los casos. Los estudios en humanos son ejemplos perfectos donde el sesgo puede surgir sin querer. Aquí

reconsideramos un estudio donde se utilizó un nuevo fármaco para tratar a pacientes con ataques cardíacos. En particular, los investigadores querían saber si el fármaco reducía las muertes en los pacientes.

Estos investigadores diseñaron un experimento aleatorizado porque querían sacar conclusiones causales sobre el efecto del fármaco. Los voluntarios del estudio[28] fueron colocados aleatoriamente en dos grupos de estudio. Un grupo, el grupo de tratamiento, recibió el fármaco. El otro grupo, llamado grupo de control, no recibió ningún tratamiento farmacológico.

[27]Este es un concepto diferente al de un grupo de control, que discutimos en el segundo principio y en la Sección 1.4.2. [28]Los sujetos humanos a menudo se llaman pacientes, voluntarios o participantes del estudio.

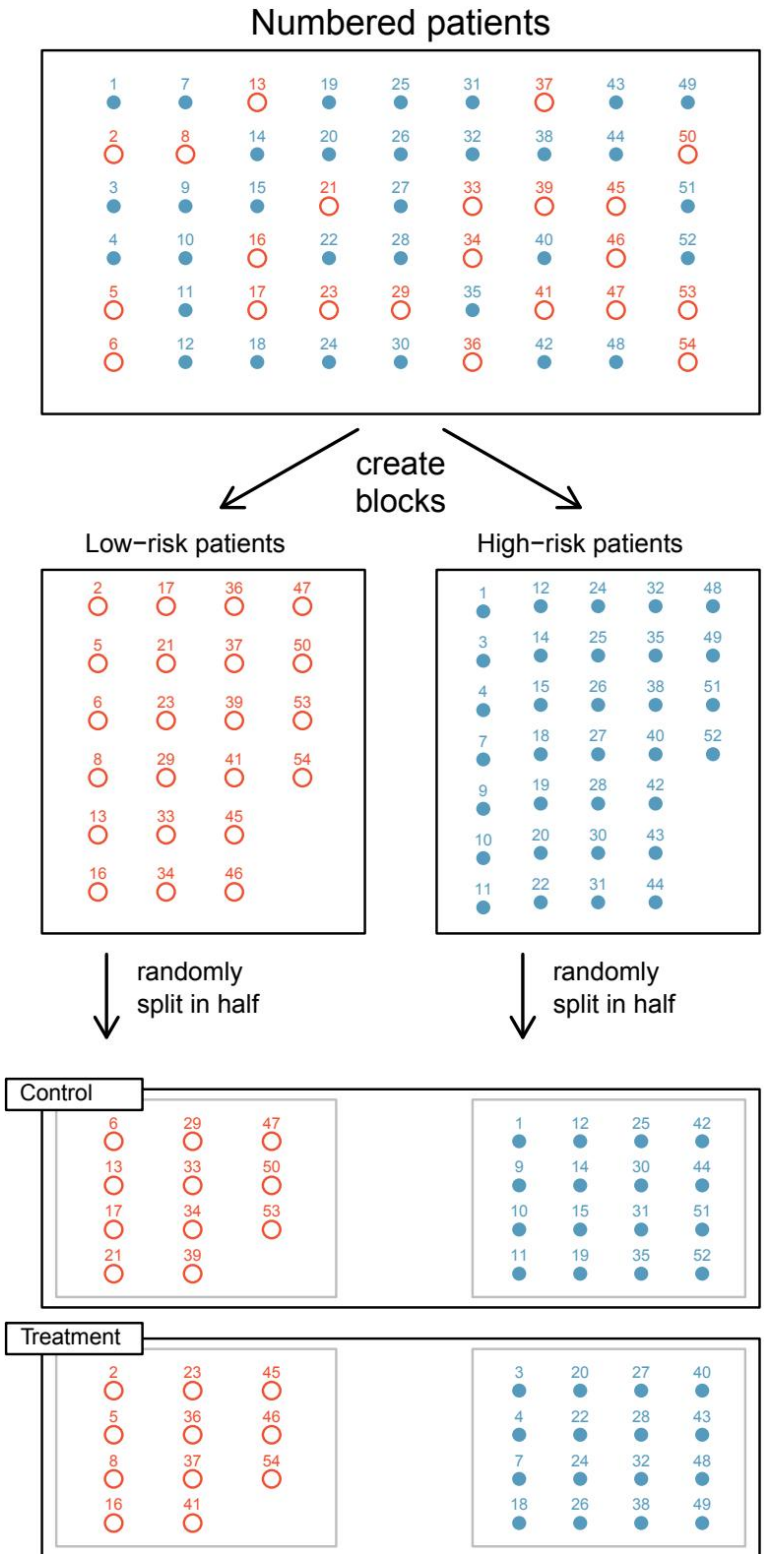

Numbered patients
create blocks
Low−risk patients
High−risk patients
randomly split in half
randomly split in half
Control
Treatment

Figura 1.16: Bloqueo usando una variable que representa el riesgo del paciente. Los pacientes se dividen primero en bloques de bajo riesgo y alto riesgo, luego cada bloque se separa uniformemente en los grupos de tratamiento utilizando la aleatorización. Esta estrategia asegura una representación igual de pacientes en cada grupo de tratamiento tanto de las categorías de bajo riesgo como de alto riesgo.

Ponte en el lugar de una persona en el estudio. Si estás en el grupo de tratamiento, te dan un nuevo fármaco elegante que esperas que te ayude. Por otro lado, una persona en el otro grupo no recibe el fármaco y se sienta ociosamente, esperando que su participación no aumente su riesgo de muerte. Estas perspectivas sugieren que en realidad hay dos efectos: el de interés es la efectividad del fármaco, y el segundo es un efecto emocional que es difícil de cuantificar.

Los investigadores no suelen estar interesados en el efecto emocional, que podría sesgar el estudio. Para evitar este problema, los investigadores no quieren que los pacientes sepan en qué grupo están. Cuando los investigadores mantienen a los pacientes desinformados sobre su tratamiento, se dice que el estudio es ciego. Pero hay un problema: si un paciente no recibe un tratamiento, sabrá que está en el grupo de control. La solución a este problema es dar tratamientos falsos a los pacientes en el grupo de control. Un tratamiento falso se llama placebo, y un placebo efectivo es la clave para hacer un estudio verdaderamente ciego. Un ejemplo clásico de un placebo es una píldora de azúcar que está hecha para parecerse a la píldora de tratamiento real. A menudo, un placebo resulta en una mejora leve pero real en los pacientes. Este efecto se ha denominado el efecto placebo.

Los pacientes no son los únicos que deben estar cegados: los médicos e investigadores pueden sesgar accidentalmente un estudio. Cuando un médico sabe que a un paciente se le ha dado el tratamiento real, podría inadvertidamente darle a ese paciente más atención o cuidado que a un paciente que sabe que está tomando el placebo. Para protegerse contra este sesgo, que de nuevo se ha encontrado que tiene un efecto medible en algunos casos, la mayoría de los estudios modernos emplean una configuración doble ciego donde los médicos o investigadores que interactúan con los pacientes son, al igual que los pacientes, desconocedores de quién está o no recibiendo el tratamiento.29

PRÁCTICA GUIADA 1.16

Vuelve al estudio de la Sección 1.1 donde los investigadores estaban probando si los stents eran eficaces para reducir los accidentes cerebrovasculares en pacientes de riesgo. ¿Es esto un experimento? ¿Estaba cegado el estudio? ¿Fue doble ciego?30

PRÁCTICA GUIADA 1.17

Para el estudio de la Sección 1.1, ¿podrían los investigadores haber empleado un placebo? Si es así, ¿cómo habría sido ese placebo?31

Es posible que tengas muchas preguntas sobre la ética de las cirugías simuladas para crear un placebo después de leer la Práctica Guiada 1.17. Estas preguntas pueden incluso haber surgido en tu mente cuando en el contexto general del experimento, donde un tratamiento posiblemente útil fue retenido de los individuos en el grupo de control; la principal diferencia es que una cirugía simulada tiende a crear un riesgo adicional, mientras que la retención de un tratamiento sólo mantiene el riesgo de una persona.

Siempre hay múltiples puntos de vista de los experimentos y placebos, y rara vez es obvio cuál es éticamente "correcto". Por ejemplo, ¿es ético usar una cirugía simulada cuando crea un riesgo para el paciente? Sin embargo, si no usamos cirugías simuladas, podemos promover el uso de un tratamiento costoso que no tiene ningún efecto real; si esto sucede, el dinero y otros recursos se desviarán de otros tratamientos que se sabe que son útiles. En última instancia, esta es una situación difícil donde no podemos proteger perfectamente tanto a los pacientes que se han ofrecido voluntariamente para el estudio como a los pacientes que pueden beneficiarse (o no) del tratamiento en el futuro.

29Siempre hay algunos investigadores involucrados en el estudio que sí saben qué pacientes están recibiendo qué tratamiento. Sin embargo, no interactúan con los pacientes del estudio y no les dicen a los profesionales de la salud cegados quién está recibiendo qué tratamiento.

30Los investigadores asignaron a los pacientes a sus grupos de tratamiento, por lo que este estudio fue un experimento. Sin embargo, los pacientes podían distinguir qué tratamiento recibieron, por lo que este estudio no fue ciego. El estudio no pudo ser doble ciego ya que no fue ciego.

31En última instancia, ¿podemos hacer que los pacientes piensen que fueron tratados por una cirugía? De hecho, podemos, y algunos experimentos utilizan lo que se llama una cirugía simulada. En una cirugía simulada, el paciente sí se somete a una cirugía, pero el paciente no recibe el tratamiento completo, aunque todavía obtendrá un efecto placebo.

Ejercicios

1.29 Luz y rendimiento en exámenes. Se diseña un estudio para probar el efecto del nivel de luz en el rendimiento en exámenes de los estudiantes. El investigador cree que los niveles de luz podrían tener diferentes efectos en hombres y mujeres, por lo que quiere asegurarse de que ambos estén representados por igual en cada tratamiento. Los tratamientos son iluminación fluorescente cenital, iluminación amarilla cenital y sin iluminación cenital (solo lámparas de escritorio).

- (a) ¿Cuál es la variable de respuesta?
- (b) ¿Cuál es la variable explicativa? ¿Cuáles son sus niveles?
- (c) ¿Cuál es la variable de bloqueo? ¿Cuáles son sus niveles?

1.30 Suplementos vitamínicos. Para evaluar la eficacia de tomar grandes dosis de vitamina C para reducir la duración del resfriado común, los investigadores reclutaron a 400 voluntarios sanos entre el personal y los estudiantes de una universidad. A una cuarta parte de los pacientes se les asignó un placebo, y el resto se dividió equitativamente entre 1 g de vitamina C, 3 g de vitamina C o 3 g de vitamina C más aditivos para tomar al inicio de un resfriado durante los dos días siguientes. Todas las pastillas tenían una apariencia y un embalaje idénticos. Las enfermeras que entregaron las píldoras recetadas a los pacientes sabían qué paciente recibió qué tratamiento, pero los investigadores que evaluaron a los pacientes cuando estaban enfermos no lo sabían. No se observaron diferencias significativas en ninguna medida de la duración o gravedad del resfriado entre los cuatro grupos, y el grupo placebo tuvo la duración más corta de los síntomas.32

- (a) ¿Fue esto un experimento o un estudio observacional? ¿Por qué?
- (b) ¿Cuáles son las variables explicativas y de respuesta en este estudio?
- (c) ¿Se cegó a los pacientes a su tratamiento?
- (d) ¿Fue este estudio doble ciego?
- (e) En última instancia, los participantes pueden elegir si usar o no las píldoras que se les recetaron. Podríamos esperar que no todos se adhieran y tomen sus píldoras. ¿Introduce esto una variable de confusión en el estudio? Explica tu razonamiento.

1.31 Luz, ruido y rendimiento en exámenes. Se diseña un estudio para probar el efecto del nivel de luz y el nivel de ruido en el rendimiento en exámenes de los estudiantes. El investigador cree que los niveles de luz y ruido podrían tener diferentes efectos en hombres y mujeres, por lo que quiere asegurarse de que ambos estén representados por igual en cada tratamiento. Los tratamientos de luz considerados son iluminación fluorescente cenital, iluminación amarilla cenital y sin iluminación cenital (solo lámparas de escritorio). Los tratamientos de ruido considerados son sin ruido, ruido de construcción y ruido de conversación humana.

- (a) ¿Qué tipo de estudio es este?
- (b) ¿Cuántos factores se consideran en este estudio? Identifíquelos y describa sus niveles.
- (c) ¿Cuál es el papel de la variable sexo en este estudio?

1.32 Música y aprendizaje. Le gustaría realizar un experimento en clase para ver si los estudiantes aprenden mejor si estudian sin música, con música sin letra (instrumental) o con música con letra. Describa brevemente un diseño para este estudio.

1.33 Preferencia por los refrescos. Le gustaría realizar un experimento en clase para ver si sus compañeros de clase prefieren el sabor de la Coca-Cola normal o la Coca-Cola Light. Describa brevemente un diseño para este estudio.

1.34 Ejercicio y salud mental. Un investigador está interesado en los efectos del ejercicio en la salud mental y propone el siguiente estudio: Utilizar un muestreo aleatorio estratificado para asegurar proporciones representativas de personas de 18-30, 31-40 y 41-55 años de la población. A continuación, asignar aleatoriamente a la mitad de los sujetos de cada grupo de edad a hacer

ejercicio dos veces por semana, e instruir al resto para que no hagan ejercicio. Realizar un examen de salud mental al principio y al final del estudio, y comparar los resultados.

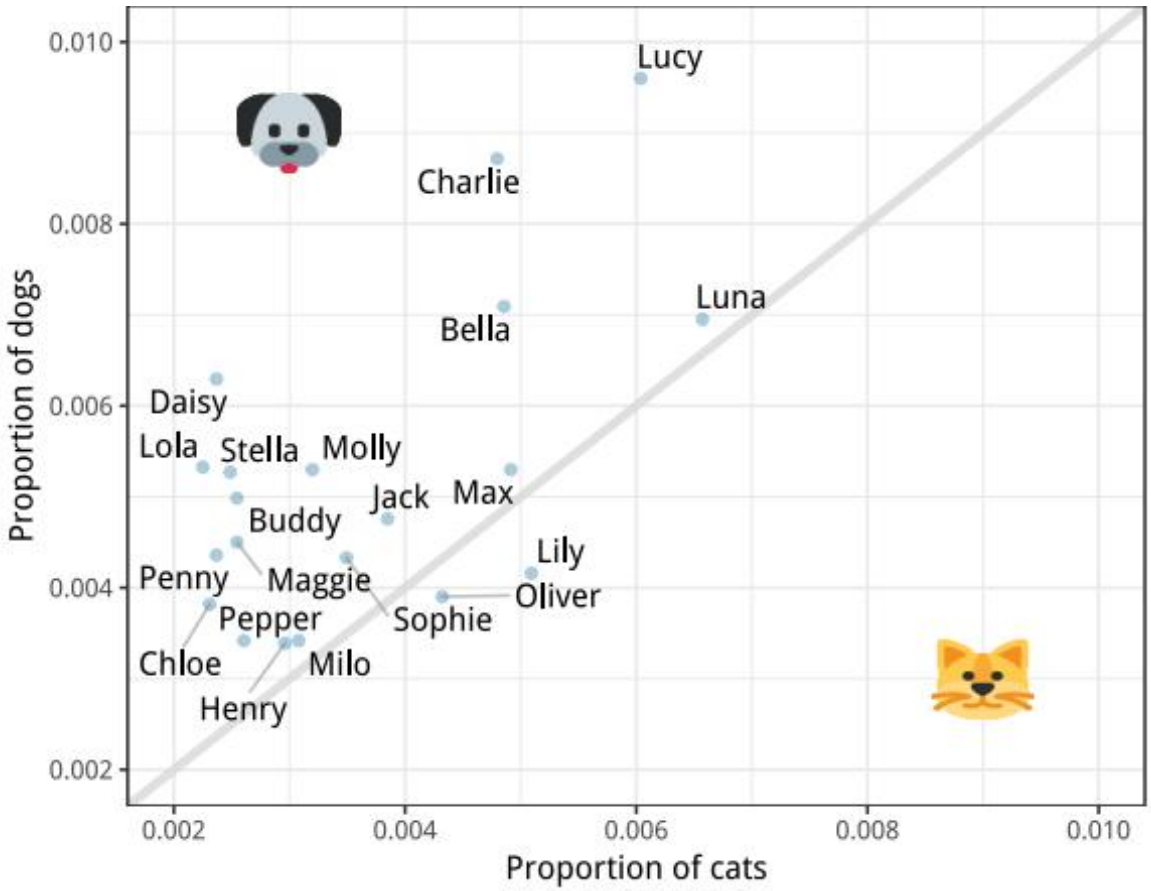
- (a) ¿Qué tipo de estudio es este?
- (b) ¿Cuáles son los grupos de tratamiento y control en este estudio?
- (c) ¿Este estudio utiliza el bloqueo? Si es así, ¿cuál es la variable de bloqueo?
- (d) ¿Este estudio utiliza el cegamiento?
- (e) Comente si los resultados del estudio pueden utilizarse para establecer una relación causal entre el ejercicio y la salud mental, e indique si las conclusiones pueden generalizarse a la población en general.
- (f) Supongamos que se le encarga la tarea de determinar si este estudio propuesto debe recibir financiación. ¿Tendría alguna reserva sobre la propuesta de estudio?

32C. Audera et al. "Mega-dose vitamin C in treatment of the common cold: a randomised controlled trial". En: Medical Journal of Australia 175.7 (2001), pp. 359–362.

Ejercicios del Capítulo

1.35 Nombres de mascotas. La ciudad de Seattle, WA tiene un portal de datos abiertos que incluye mascotas registradas en la ciudad. Para cada mascota registrada, tenemos información sobre el nombre y la especie de la mascota. La siguiente visualización grafica la proporción de perros con un nombre dado versus la proporción de gatos con el mismo nombre. Se muestran los 20 nombres de gatos y perros más comunes. La línea diagonal en el gráfico es la línea x = y; si un nombre apareciera en esta línea, la popularidad del nombre sería exactamente la misma para perros y gatos.

- (a) ¿Estos datos se recopilaron como parte de un experimento o de un estudio observacional?
- (b) ¿Cuál es el nombre de perro más común? ¿Cuál es el nombre de gato más común?
- (c) ¿Qué nombres son más comunes para gatos que para perros?
- (d) ¿La relación entre las dos variables es positiva o negativa? ¿Qué significa esto en el contexto de los datos?



1.36 Estresado, Parte II. En un estudio que evalúa la relación entre el estrés y los calambres musculares, la mitad de los sujetos son asignados aleatoriamente a ser expuestos a un aumento del estrés al ser colocados en un ascensor que cae rápidamente y se detiene abruptamente y la otra mitad se deja sin estrés o con estrés basal.

- (a) ¿Qué tipo de estudio es este?
- (b) ¿Se puede utilizar este estudio para concluir una relación causal entre el aumento del estrés y los calambres musculares?

1.37 Semillas de chía y pérdida de peso. Chia Pets – esas figuritas de terracota a las que les brota un cabello verde y difuso – hicieron de la planta de chía un nombre familiar. Pero la chía ha ganado una reputación completamente nueva como suplemento dietético. En un estudio de 2009, un equipo de investigadores reclutó a 38 hombres y los dividió aleatoriamente en dos grupos: tratamiento o control. También reclutaron a 38 mujeres y asignaron aleatoriamente a la mitad de estas participantes al grupo de tratamiento y a la otra mitad al grupo de control. Un grupo recibió 25 gramos de semillas de chía dos veces al día y el otro recibió un placebo. Los sujetos se ofrecieron voluntariamente a participar en el estudio. Después de 12 semanas, los

científicos no encontraron diferencias significativas entre los grupos en el apetito o la pérdida de peso.33

- (a) ¿Qué tipo de estudio es este?
- (b) ¿Cuáles son los tratamientos experimentales y de control en este estudio?
- (c) ¿Se ha utilizado el bloqueo en este estudio? Si es así, ¿cuál es la variable de bloqueo?
- (d) ¿Se ha utilizado el cegamiento en este estudio?
- (e) Comente si podemos o no hacer una declaración causal e indique si podemos o no generalizar la conclusión a la población en general.

1.38 Encuesta del consejo municipal. Un consejo municipal ha solicitado que se realice una encuesta de hogares en una zona suburbana de su ciudad. La zona se divide en muchos barrios distintos y únicos, algunos con casas grandes, otros solo con apartamentos y otros con una mezcla diversa de estructuras de vivienda. Para cada parte a continuación, identifique los métodos de muestreo descritos y describa los pros y los contras estadísticos del método en el contexto de la ciudad.

- (a) Muestrear aleatoriamente 200 hogares de la ciudad.
- (b) Dividir la ciudad en 20 barrios y muestrear 10 hogares de cada barrio.
- (c) Dividir la ciudad en 20 barrios, muestrear aleatoriamente 3 barrios y luego muestrear todos los hogares de esos 3 barrios.
- (d) Dividir la ciudad en 20 barrios, muestrear aleatoriamente 8 barrios y luego muestrear aleatoriamente 50 hogares de esos barrios.
- (e) Muestrear los 200 hogares más cercanos a las oficinas del consejo municipal.

33D.C. Nieman et al. "La semilla de chía no promueve la pérdida de peso ni altera los factores de riesgo de enfermedad en adultos con sobrepeso". En: Nutrition Research 29.6 (2009), pp. 414–418.

1.4. EXPERIMENTOS 37

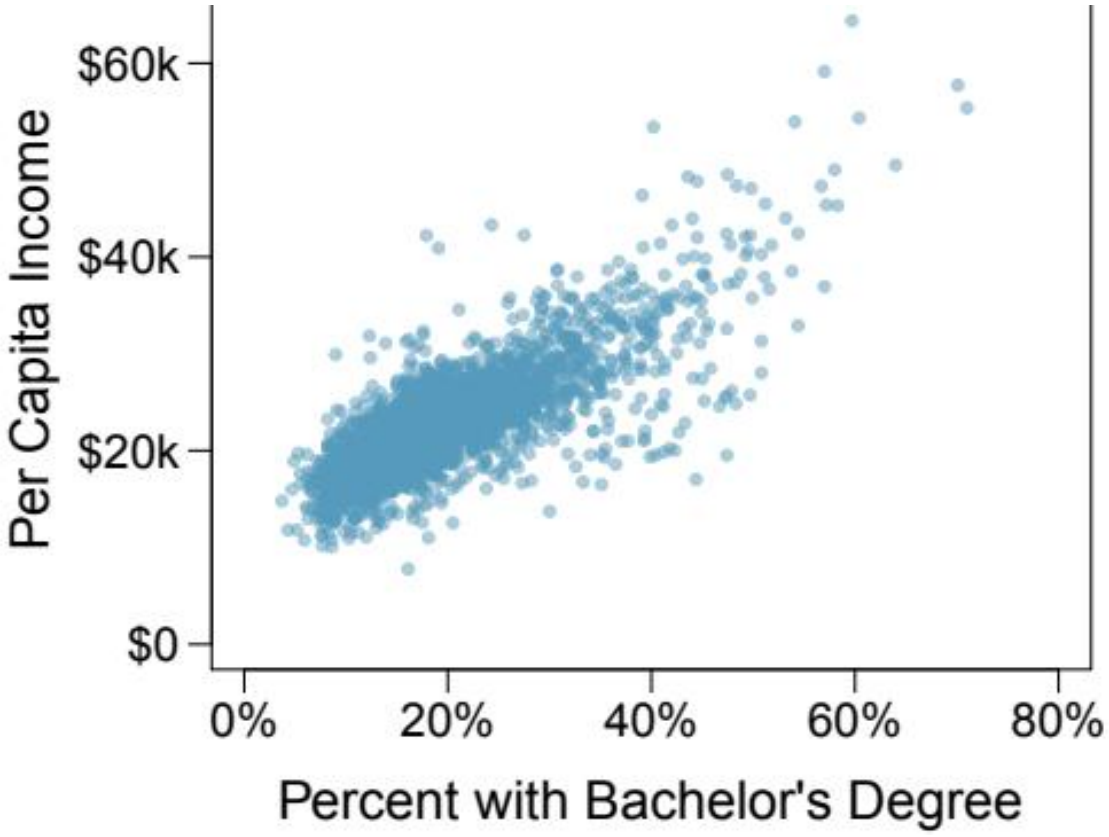
1.39 Razonamiento defectuoso. Identifique la(s) falla(s) en el razonamiento en los siguientes escenarios. Explique qué deberían haber hecho de manera diferente los individuos en el estudio si querían sacar conclusiones tan fuertes.

- (a) A los estudiantes de una escuela primaria se les entrega un cuestionario que se les pide que devuelvan después de que sus padres lo hayan completado. Una de las preguntas formuladas es: "¿Considera que su horario de trabajo le dificulta pasar tiempo con sus hijos después de la escuela?" De los padres que respondieron, el 85% dijo "no". Con base en estos resultados, los funcionarios de la escuela concluyen que una gran mayoría de los padres no tienen dificultad para pasar tiempo con sus hijos después de la escuela.

- (b) Se realiza una encuesta en una muestra aleatoria simple de 1,000 mujeres que recientemente dieron a luz, preguntándoles si fumaron o no durante el embarazo. Una encuesta de seguimiento que pregunta si los niños tienen problemas respiratorios se lleva a cabo 3 años después. Sin embargo, solo se localiza a 567 de estas mujeres en la misma dirección. El investigador informa que estas 567 mujeres son representativas de todas las madres.
- (c) Un ortopedista administra un cuestionario a 30 de sus pacientes que no tienen ningún problema articular y encuentra que 20 de ellos corren con regularidad. Concluye que correr disminuye el riesgo de problemas articulares.

1.40 Ingresos y educación en los condados de EE. UU. El diagrama de dispersión a continuación muestra la relación entre el ingreso per cápita (en miles de dólares) y el porcentaje de población con una licenciatura en 3,143 condados en los EE. UU. en 2010.

- (a) ¿Cuáles son las variables explicativa y de respuesta?
- (b) Describe la relación entre las dos variables. Asegúrese de discutir las observaciones inusuales, si las hay.
- (c) ¿Podemos concluir que tener una licenciatura aumenta los ingresos de uno?



1.41 ¿Comer mejor, sentirse mejor? En un estudio de salud pública sobre los efectos del consumo de frutas y verduras en el bienestar psicológico en adultos jóvenes, los participantes fueron asignados aleatoriamente a tres grupos: (1) dietas habituales, (2) una intervención ecológica momentánea que involucraba recordatorios por mensaje de texto para aumentar su consumo de frutas y verduras más un vale para comprarlas, o (3) una intervención de frutas y verduras en la que se les daba a los participantes dos porciones diarias adicionales de frutas y verduras frescas para consumir además de su dieta normal. Se les pidió a los participantes que realizaran una encuesta nocturna en sus teléfonos inteligentes. Los participantes fueron estudiantes voluntarios de la Universidad de Otago, Nueva Zelanda. Al final del estudio de 14 días, solo los participantes en el tercer grupo mostraron mejoras en su bienestar psicológico durante los 14 días en relación con los otros grupos.34

- (a) ¿Qué tipo de estudio es este?
- (b) Identifique las variables explicativas y de respuesta.
- (c) Comente si los resultados del estudio se pueden generalizar a la población.
- (d) Comente si los resultados del estudio se pueden utilizar para establecer relaciones causales.
- (e) Un artículo periodístico que informa sobre el estudio afirma: "Los resultados de este estudio demuestran que dar a los adultos jóvenes frutas y verduras frescas para comer puede tener beneficios psicológicos, incluso durante un breve período de tiempo". ¿Cómo sugeriría revisar esta declaración para que pueda ser respaldada por el estudio?

1.42 Pantallas, adolescentes y bienestar psicológico. En un estudio de tres grandes conjuntos de datos representativos a nivel nacional de Irlanda, Estados Unidos y el Reino Unido (n = 17,247), se les pidió a los adolescentes entre las edades de 12 a 15 años que llevaran un diario de su tiempo frente a la pantalla y respondieran preguntas sobre cómo se sentían o actuaban. Las respuestas a estas preguntas se utilizaron luego para calcular una puntuación de bienestar psicológico. Se recopilaron datos adicionales y se incluyeron en el análisis, como el sexo y la edad de cada niño, y la educación, el origen étnico, la angustia psicológica y el empleo de la madre. El estudio concluyó que hay poca evidencia clara de que el tiempo frente a la pantalla disminuya el bienestar de los adolescentes.35

- (a) ¿Qué tipo de estudio es este?
- (b) Identifique las variables explicativas.
- (c) Identifique la variable de respuesta.
- (d) Comente si los resultados del estudio se pueden generalizar a la población y por qué.
- (e) Comente si los resultados del estudio se pueden utilizar para establecer relaciones causales.

1.43 Stanford Open Policing. El proyecto Stanford Open Policing recopila, analiza y publica registros de paradas de tráfico por parte de las agencias de aplicación de la ley en los Estados Unidos. Su objetivo es ayudar a los investigadores, periodistas y formuladores de políticas a investigar y mejorar las interacciones entre la policía y el público.36 El siguiente es

un extracto de una tabla resumen creada a partir de los datos recopilados como parte de este proyecto.

Condado	Estado	Raza del conductor	No. de paradas por año	coches registrados	% de conductores arrestados
Apaice County	Arizona	Negra	266	0.08	0.02
Apaice County	Arizona	Hispana	1008	0.05	0.02
Apaice County	Arizona	Blanca	6322	0.02	0.01
Cochise County	Arizona	Negra	1169	0.05	0.01
Cochise County	Arizona	Hispana	9453	0.04	0.01
Cochise County	Arizona	Blanca	10826	0.02	0.01
. . .	. . .	. . .	. . .	. . .	. . .
Wood County	Wisconsin	Negra	16	0.24	0.10
Wood County	Wisconsin	Hispana	27	0.04	0.03
Wood County	Wisconsin	Blanca	1157	0.03	0.03

- (a) ¿Qué variables se recopilaron en cada parada de tráfico individual para crear la tabla resumen anterior?
- (b) Indique si cada variable es numérica o categórica. Si es numérica, indique si es continua o discreta. Si es categórica, indique si es ordinal o no.
- (c) Supongamos que queremos evaluar si las tasas de registro de vehículos son diferentes para los conductores de diferentes razas. En este análisis, ¿qué variable sería la variable de respuesta y qué variable sería la variable explicativa?

1.44 Lanzamientos espaciales. La siguiente tabla resumen muestra el número de lanzamientos espaciales en los EE. UU. por el tipo de agencia de lanzamiento y el resultado del lanzamiento (éxito o fracaso).37

	1957 - 1999		2000 - 2018	
	Fracaso	Éxito	Fracaso	Éxito
Privada	13	295	10	562

		1957 - 1999	2000 - 2018	
Estatal	281	3751	33	711
Startup	-	-	5	65

- (a) ¿Qué variables se recopilaron en cada lanzamiento para crear la tabla resumen anterior?
- (b) Indique si cada variable es numérica o categórica. Si es numérica, indique si es continua o discreta. Si es categórica, indique si es ordinal o no.
- (c) Supongamos que queremos estudiar cómo varía la tasa de éxito de los lanzamientos entre las agencias de lanzamiento y con el tiempo. En este análisis, ¿qué variable sería la variable de respuesta y qué variable sería la variable explicativa?

37Base de datos de vehículos de lanzamiento JSR, Una lista completa de lanzamientos espaciales suborbitales, edición del 10 de febrero de 2019.

35Amy Orben y AK Baukney-Przybylski. "Pantallas, adolescentes y bienestar psicológico: evidencia de tres estudios de diario de uso del tiempo". En: Psychological Science (2018).

36Emma Pierson et al. "Un análisis a gran escala de las disparidades raciales en las paradas policiales en los Estados Unidos". En: arXiv preprint arXiv:1706.05678 (2017).

Capítulo 2

39

Resumiendo datos

- **2.1 Examinando datos numéricos**
- **2.2 Considerando datos categóricos**
- **2.3 Estudio de caso: vacuna contra la malaria**

Este capítulo se centra en la mecánica y la construcción de estadísticas resumidas y gráficos. Utilizamos software estadístico para generar los resúmenes y gráficos presentados en este capítulo y libro. Sin embargo, dado que esta podría ser su primera exposición a estos conceptos, nos tomamos nuestro tiempo en este capítulo para detallar cómo crearlos. El dominio del contenido presentado en este capítulo será crucial para comprender los métodos y técnicas introducidos en el resto del libro.

Para videos, diapositivas y otros recursos, visite www.openintro.org/os

2.1 Examinando datos numéricos

En esta sección exploraremos técnicas para resumir variables numéricas. Por ejemplo, considere la variable de monto del préstamo del conjunto de datos loan50, que representa el tamaño del préstamo para los 50 préstamos en el conjunto de datos. Esta variable es numérica ya que podemos discutir sensatamente la diferencia numérica del tamaño de dos préstamos. Por otro lado, los códigos de área y los códigos postales no son numéricos, sino que son variables categóricas.

A lo largo de esta sección y la siguiente, aplicaremos estos métodos utilizando los conjuntos de datos loan50 y county, que se presentaron en la Sección 1.2. Si desea revisar las variables de cualquiera de los conjuntos de datos, consulte las Figuras 1.3 y 1.5.

2.1.1 Diagramas de Dispersión para Datos Pareados

Un diagrama de dispersión proporciona una vista caso por caso de los datos para dos variables numéricas. En la Figura 1.8 en la página 16, se utilizó un diagrama de dispersión para examinar la tasa de propiedad de vivienda en contra de la fracción de unidades de vivienda que formaban parte de propiedades de unidades múltiples (por ejemplo, apartamentos) en el conjunto de datos del condado. Otro diagrama de dispersión se muestra en la Figura 2.1, comparando el ingreso total de un prestatario (ingreso total) y la cantidad que pidió prestada (monto del préstamo) para el conjunto de datos loan50. En cualquier diagrama de dispersión, cada punto representa un solo caso. Dado que hay 50 casos en loan50, hay 50 puntos en la Figura 2.1.

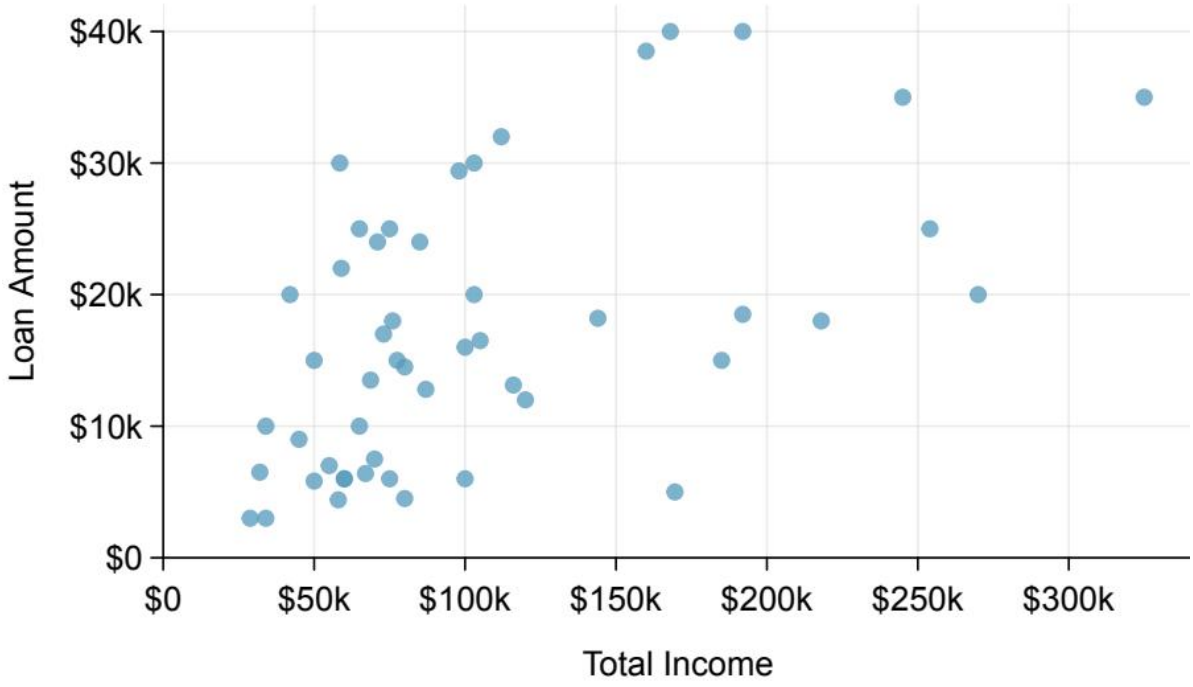


Figura 2.1: Un diagrama de dispersión del ingreso total versus el monto del préstamo para el conjunto de datos loan50.

Mirando la Figura 2.1, vemos que hay muchos prestatarios con un ingreso por debajo de $100,000 en el lado izquierdo del gráfico, mientras que hay un puñado de prestatarios con ingresos superiores a $250,000.

EJEMPLO 2.1

La Figura 2.2 muestra un gráfico del ingreso familiar mediano contra la tasa de pobreza para 3,142 condados. ¿Qué se puede decir sobre la relación entre estas variables?

La relación es evidentemente no lineal, como se destaca con la línea discontinua. Esto es diferente de los diagramas de dispersión anteriores que hemos visto, que muestran relaciones que no muestran mucha, si es que alguna, curvatura en la tendencia.

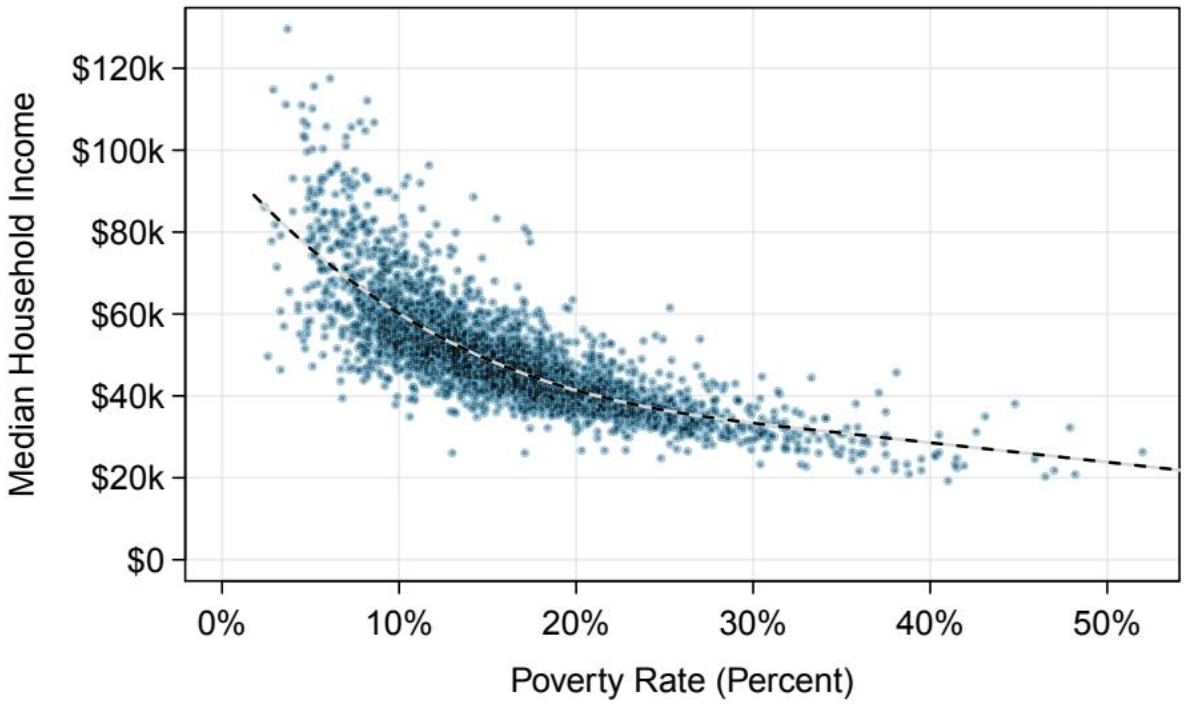


Figura 2.2: Un diagrama de dispersión del ingreso familiar mediano contra la tasa de pobreza para el conjunto de datos del condado. También se ha ajustado un modelo estadístico a los datos y se muestra como una línea discontinua.

¿Qué revelan los diagramas de dispersión sobre los datos y cómo son útiles?1

PRÁCTICA GUIADA 2.3

Describe dos variables que tendrían una asociación en forma de herradura en un diagrama de dispersión (o).2

2.1.2 Diagramas de puntos y la media

A veces, dos variables son demasiadas: solo una variable puede ser de interés. En estos casos, un diagrama de puntos proporciona la visualización más básica. Un diagrama de puntos es un diagrama de dispersión de una variable; un ejemplo que utiliza la tasa de interés de 50 préstamos se muestra en la Figura 2.3. Una versión apilada de este diagrama de puntos se muestra en la Figura 2.4.

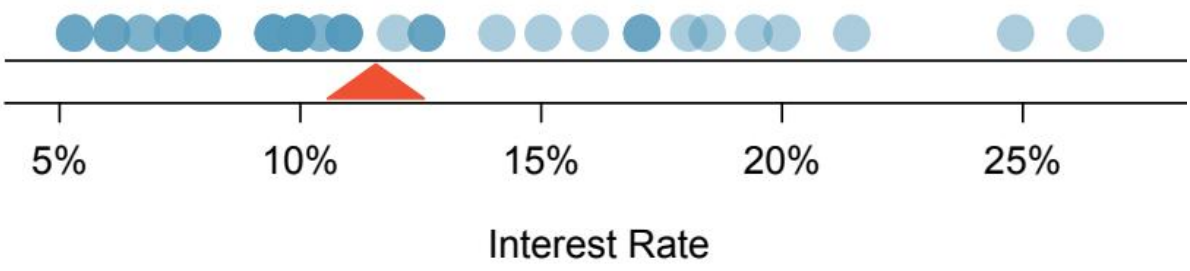


Figura 2.3: Un diagrama de puntos de la tasa de interés para el conjunto de datos loan50. La media de la distribución se muestra como un triángulo rojo.

1Las respuestas pueden variar. Los diagramas de dispersión son útiles para detectar rápidamente asociaciones que relacionan variables, ya sea que esas asociaciones vengan en forma de tendencias simples o que esas relaciones sean más complejas.

2Considere el caso en el que su eje vertical representa algo "bueno" y su eje horizontal representa algo que solo es bueno con moderación. La salud y el consumo de agua se ajustan a esta descripción: necesitamos algo de agua para sobrevivir, pero consumimos demasiado y se vuelve tóxico y puede matar a una persona.

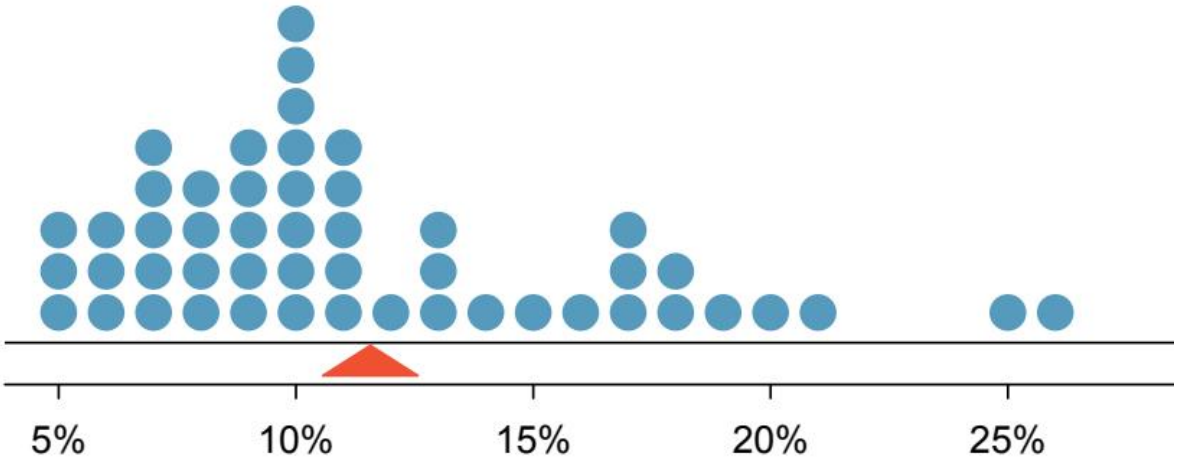


Figura 2.4: Un diagrama de puntos apilado de la tasa de interés para el conjunto de datos loan50. Las tasas se han redondeado al porcentaje más cercano en este diagrama, y la media de la distribución se muestra como un triángulo rojo.

La media, a menudo llamada el promedio, es una forma común de medir el centro de una distribución de datos. Para calcular la tasa de interés media, sumamos todas las tasas de interés y dividimos por el número de observaciones:

$$\bar{x} = \frac{10.90\% + 9.92\% + 26.30\% + \cdots + 6.08\%}{50} = 11.57\%$$

La media muestral a menudo se etiqueta como ¯x. La letra x se usa como un marcador de posición genérico para la variable de interés, la tasa de interés, y la barra sobre la x comunica que estamos viendo la tasa de interés promedio, que para estos 50 préstamos fue del 11.57%. Es útil pensar en la media como el punto de equilibrio de la distribución, y se muestra como un triángulo en las Figuras 2.3 y 2.4.

MEDIA

La media muestral puede calcularse como la suma de los valores observados dividida entre el número de observaciones:

$$\bar{x} = \frac{x\_1 + x\_2 + \cdots + x\_n}{n}$$

donde x1, x2, . . . , xn representan los n valores observados.

PRÁCTICA GUIADA 2.4

Examina la ecuación para la media. ¿A qué corresponde x1? ¿Y x2? ¿Puedes inferir un significado general a lo que xi podría representar?3

PRÁCTICA GUIADA 2.5

¿Cuál era n en esta muestra de préstamos?4

El conjunto de datos loan50 representa una muestra de una población más grande de préstamos realizados a través de Lending Club. Podríamos calcular una media para esta población de la misma manera que la media muestral. Sin embargo, la media poblacional tiene una etiqueta especial: µ. El símbolo µ es la letra griega mu y representa el promedio de todas las observaciones en la población. A veces se usa un subíndice, como x, para representar a qué variable se refiere la media poblacional, p. ej. µx. A menudo, es demasiado costoso medir la media poblacional con precisión, por lo que a menudo estimamos µ usando la media muestral, ¯x.

3x1 corresponde a la tasa de interés para el primer préstamo en la muestra (10.90%), x2 a la tasa de interés del segundo préstamo (9.92%) y xi corresponde a la tasa de interés para el i-ésimo préstamo en el conjunto de datos. Por ejemplo, si i = 4, entonces estamos examinando x4, que se refiere a la cuarta observación en el conjunto de datos.

4El tamaño de la muestra fue n = 50.

EJEMPLO 2.6

La tasa de interés promedio en todos los préstamos en la población se puede estimar utilizando los datos de la muestra. Según la muestra de 50 préstamos, ¿cuál sería una estimación razonable de µx, la tasa de interés media para todos los préstamos en el conjunto de datos completo?

La media muestral, 11.57%, proporciona una estimación aproximada de µx. Si bien no es perfecto, esta es nuestra mejor conjetura de la tasa de interés promedio de todos los préstamos en la población bajo estudio.

En el Capítulo 5 y más adelante, desarrollaremos herramientas para caracterizar la precisión de las estimaciones puntuales como la media muestral. Como habrás adivinado, las estimaciones puntuales basadas en muestras más grandes tienden a ser más precisas que las basadas en muestras más pequeñas.

EJEMPLO 2.7

La media es útil porque nos permite reescalar o estandarizar una métrica en algo más fácil de interpretar y comparar. Proporciona 2 ejemplos donde la media es útil para hacer comparaciones.

1. Nos gustaría entender si un nuevo fármaco es más eficaz para tratar los ataques de asma que el fármaco estándar. Se establece un ensayo con 1500 adultos, donde 500 reciben el nuevo fármaco y 1000 reciben un fármaco estándar en el grupo de control:

	Nuevo fármaco	Fármaco estándar
Número de pacientes	500	1000
Ataques totales de asma	200	300

Comparar los recuentos brutos de 200 a 300 ataques de asma haría parecer que el nuevo fármaco es mejor, pero esto es un artefacto del tamaño desigual de los grupos. En cambio, deberíamos observar el número medio de ataques de asma por paciente en cada grupo:

Nuevo fármaco: 200/500 = 0.4 Fármaco estándar: 300/1000 = 0.3

El fármaco estándar tiene un número medio de ataques de asma por paciente menor que la media del grupo de tratamiento.

2. Emilio abrió un camión de comida el año pasado donde vende burritos, y su negocio se ha estabilizado en los últimos 3 meses. Durante ese período de 3 meses, ha ganado $11,000 mientras trabajaba 625 horas. Las ganancias medias por hora de Emilio proporcionan una estadística útil para evaluar si su empresa, al menos desde una perspectiva financiera, merece la pena:

   $11000 625 horas = $17.60 por hora

Al conocer su salario medio por hora, Emilio ha convertido sus ganancias en una unidad estándar que es más fácil de comparar con muchos otros trabajos que podría considerar.

EJEMPLO 2.8

Supongamos que queremos calcular el ingreso medio por persona en los EE. UU. Para ello, podríamos pensar en calcular la media de los ingresos per cápita en los 3142 condados del conjunto de datos del condado. ¿Cuál sería un mejor enfoque?

El conjunto de datos del condado es especial en el sentido de que cada condado representa en realidad a muchas personas individuales. Si simplemente promediáramos la variable de ingresos, estaríamos tratando a los condados con 5000 y 5,000,000 de residentes por igual en los cálculos. En cambio, deberíamos calcular el ingreso total para cada condado, sumar los totales de todos los condados y luego dividir por el número de personas en todos los condados. Si completáramos estos pasos con los datos del condado, encontraríamos que el ingreso per cápita para los EE. UU. es de $30,861. Si hubiéramos calculado la media simple del ingreso per cápita entre los condados, ¡el resultado habría sido de solo $26,093!

Este ejemplo utilizó lo que se llama una media ponderada. Para obtener más información sobre este tema, consulta el siguiente suplemento en línea sobre medias ponderadas openintro.org/d?file=stat wtd mean.

2.1.3 Histogramas y forma

Los diagramas de puntos muestran el valor exacto para cada observación. Esto es útil para conjuntos de datos pequeños, pero pueden volverse difíciles de leer con muestras más grandes. En lugar de mostrar el valor de cada observación, preferimos pensar en el valor como perteneciente a un intervalo. Por ejemplo, en el conjunto de datos loan50, creamos una tabla de recuentos para el número de préstamos con tasas de interés entre 5.0% y 7.5%, luego el número de préstamos con tasas entre 7.5% y 10.0%, y así sucesivamente. Las observaciones que caen en el límite de un intervalo (por ejemplo, 10.00%) se asignan al intervalo inferior. Esta tabulación se muestra en la Figura 2.5. Estos recuentos agrupados se representan como barras en la Figura 2.6 en lo que se llama un histograma, que se asemeja a una versión más agrupada del diagrama de puntos apilados que se muestra en la Figura 2.4.

Tasa de interés	5.0% - 7.5%	7.5% - 10.0%	10.0% - 12.5%	12.5% - 15.0%	· · ·	25.0% - 27.5%
Recuento	11	15	8	4	· · ·	1

Figura 2.5: Recuentos para los datos de la tasa de interés agrupados.

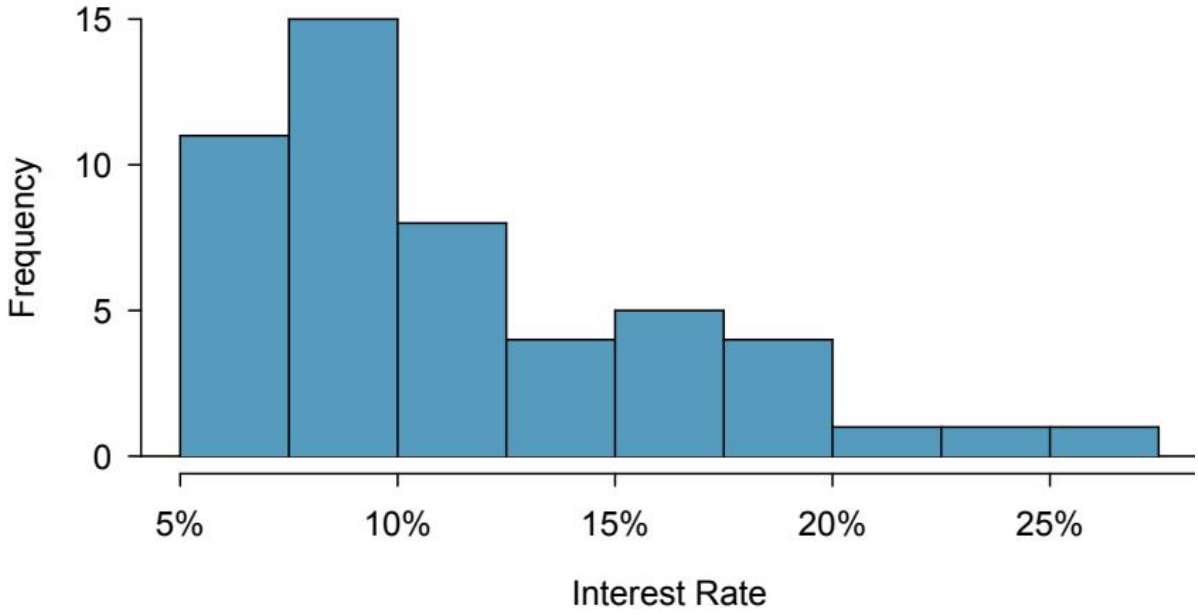


Figura 2.6: Un histograma de la tasa de interés. Esta distribución está fuertemente sesgada hacia la derecha.

Los histogramas proporcionan una vista de la densidad de los datos. Las barras más altas representan dónde los datos son relativamente más comunes. Por ejemplo, hay muchos más préstamos con tasas entre el 5% y el 10% que préstamos con tasas entre el 20% y el 25% en el conjunto de datos. Las barras facilitan ver cómo cambia la densidad de los datos en relación con la tasa de interés.

Los histogramas son especialmente convenientes para comprender la forma de la distribución de los datos. La Figura 2.6 sugiere que la mayoría de los préstamos tienen tasas por debajo del 15%, mientras que solo un puñado de préstamos tienen tasas por encima del 20%. Cuando los datos se extienden hacia la derecha de esta manera y tienen una cola derecha más larga, se dice que la forma está sesgada a la derecha. 5

Los conjuntos de datos con la característica inversa, una cola larga y delgada a la izquierda, se dice que están sesgados a la izquierda. También decimos que tal distribución tiene una cola larga a la izquierda. Los conjuntos de datos que muestran una atenuación aproximadamente igual en ambas direcciones se llaman simétricos.

COLAS LARGAS PARA IDENTIFICAR EL SESGO

Cuando los datos se extienden en una dirección, la distribución tiene una cola larga. Si una distribución tiene una cola larga a la izquierda, está sesgada a la izquierda. Si una distribución tiene una cola larga a la derecha, está sesgada a la derecha.

5Otras formas de describir los datos que están sesgados a la derecha: sesgados hacia la derecha, sesgados hacia el extremo superior o sesgados hacia el extremo positivo.

Echa un vistazo a los diagramas de puntos en las Figuras 2.3 y 2.4. ¿Puedes ver el sesgo en los datos? ¿Es más fácil ver el sesgo en este histograma o en los diagramas de puntos?6

PRÁCTICA GUIADA 2.10

Además de la media (ya que fue etiquetada), ¿qué puedes ver en los diagramas de puntos que no puedes ver en el histograma?7

Además de observar si una distribución está sesgada o es simétrica, los histogramas se pueden utilizar para identificar modos. Un modo está representado por un pico prominente en la distribución. Solo hay un pico prominente en el histograma del monto del préstamo.

Una definición de modo que a veces se enseña en las clases de matemáticas es el valor con más ocurrencias en el conjunto de datos. Sin embargo, para muchos conjuntos de datos del mundo real, es común no tener observaciones con el mismo valor en un conjunto de datos, lo que hace que esta definición no sea práctica en el análisis de datos.

La Figura 2.7 muestra histogramas que tienen uno, dos o tres picos prominentes. Tales distribuciones se llaman unimodales, bimodales y multimodales, respectivamente. Cualquier distribución con más de 2 picos prominentes se llama multimodal. Observa que había un pico prominente en la distribución unimodal con un segundo pico menos prominente que no se contó ya que solo difiere de sus intervalos vecinos en unas pocas observaciones.

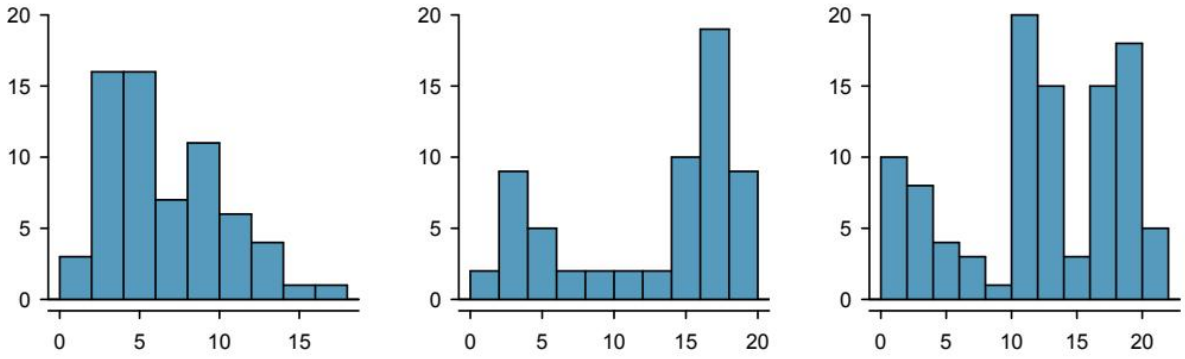


Figura 2.7: Contando solo los picos prominentes, las distribuciones son (de izquierda a derecha) unimodal, bimodal y multimodal. Ten en cuenta que hemos dicho que la gráfica de la izquierda es unimodal intencionalmente. Esto se debe a que estamos contando picos prominentes, no solo cualquier pico.

EJEMPLO 2.11

La Figura 2.6 revela solo un modo prominente en la tasa de interés. ¿Es la distribución unimodal, bimodal o multimodal?

Unimodal. Recuerda que uni significa 1 (piensa en monociclos). De manera similar, bi significa 2 (piensa en bicicletas). Esperamos que se invente un multiciclo para completar esta analogía.

PRÁCTICA GUIADA 2.12

Se tomaron medidas de altura de estudiantes jóvenes y profesores adultos en una escuela primaria K-3. ¿Cuántos modos esperaría encontrar en este conjunto de datos de altura?8

La búsqueda de modos no se trata de encontrar una respuesta clara y correcta sobre el número de modos en una distribución, razón por la cual "prominente" no está rigurosamente definido en este libro. La parte más importante de este examen es comprender mejor sus datos.

6La asimetría es visible en los tres gráficos, aunque el diagrama de puntos plano es el menos útil. El diagrama de puntos apilados y el histograma son visualizaciones útiles para identificar la asimetría.

7Las tasas de interés de los préstamos individuales.

8Podrían haber dos grupos de altura visibles en el conjunto de datos: uno de los estudiantes y otro de los adultos. Es decir, los datos son probablemente bimodales.

2.1.4 Varianza y desviación estándar

La media se introdujo como un método para describir el centro de un conjunto de datos, y la variabilidad en los datos también es importante. Aquí, presentamos dos medidas de variabilidad: la varianza y la desviación estándar. Ambas son muy útiles en el análisis de datos, aunque sus fórmulas son un poco tediosas de calcular a mano. La desviación estándar es la más fácil de comprender de las dos, y describe aproximadamente qué tan lejos está la observación típica de la media.

Llamamos a la distancia de una observación desde su media su desviación. A continuación, se muestran las desviaciones para las observaciones 1^{a}, 2^{a}, 3^{a} y 50^{a} en la variable de tasa de interés:

$$
\begin{aligned}
x\_1 - \bar{x} &= 10.90 - 11.57 = -0.67 \\
x\_2 - \bar{x} &= 9.92 - 11.57 = -1.65 \\
x\_3 - \bar{x} &= 26.30 - 11.57 = 14.73 \\
&\vdots \\
x\_50 - \bar{x} &= 6.08 - 11.57 = -5.49
\end{aligned}
$$

Si elevamos al cuadrado estas desviaciones y luego sacamos un promedio, el resultado es igual a la varianza de la muestra, denotada por s 2 :

$$
\begin{aligned}
s^2 &= \frac{(-0.67)^2 + (-1.65)^2 + (14.73)^2 + \cdots + (-5.49)^2}{50 - 1} \\
&= \frac{0.45 + 2.72 + 216.97 + \cdots + 30.14}{49} \\
&= 25.52
\end{aligned}
$$

Dividimos por n − 1, en lugar de dividir por n, al calcular la varianza de una muestra; hay algunos matices matemáticos aquí, pero el resultado final es que hacer esto hace que esta estadística sea un poco más confiable y útil.

Observe que elevar al cuadrado las desviaciones hace dos cosas. Primero, hace que los valores grandes sean relativamente mucho más grandes, como se ve al comparar (−0.67)2 , (−1.65)2 , (14.73)2 y (−5.49)2 . En segundo lugar, elimina cualquier signo negativo.

La desviación estándar se define como la raíz cuadrada de la varianza:

$$
s = \sqrt{25.52} = 5.05
$$

Si bien a menudo se omite, se puede agregar un subíndice de x a la varianza y la desviación estándar, es decir, s 2 x y sx , si es útil como recordatorio de que estas son la varianza y la desviación estándar de las observaciones representadas por x1 , x2 , …, xn .

VARIANZA Y DESVIACIÓN ESTÁNDAR

La varianza es la distancia promedio al cuadrado desde la media. La desviación estándar es la raíz cuadrada de la varianza. La desviación estándar es útil cuando se considera qué tan lejos están distribuidos los datos de la media.

La desviación estándar representa la desviación típica de las observaciones con respecto a la media. Generalmente, alrededor del 70% de los datos estarán dentro de una desviación estándar

de la media y alrededor del 95% estarán dentro de dos desviaciones estándar. Sin embargo, como se ve en las Figuras 2.8 y 2.9, estos porcentajes no son reglas estrictas.

Al igual que la media, los valores de la población para la varianza y la desviación estándar tienen símbolos especiales: 2 para la varianza y para la desviación estándar. El símbolo es la letra griega sigma.

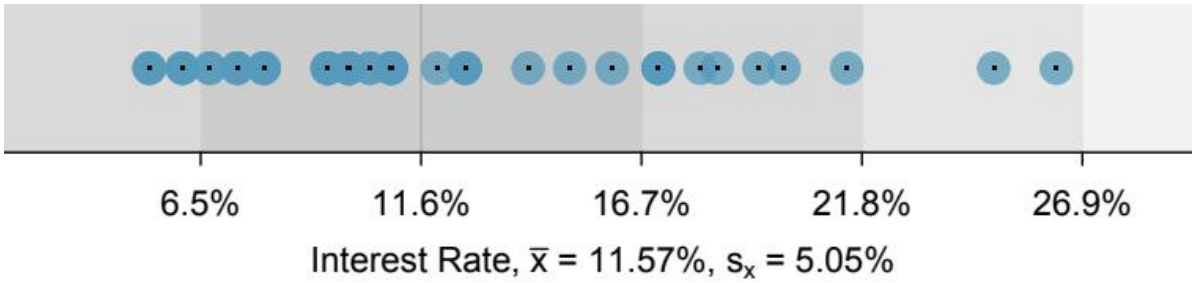


Figura 2.8: Para la variable de tasa de interés, 34 de los 50 préstamos (68%) tenían tasas de interés dentro de 1 desviación estándar de la media, y 48 de los 50 préstamos (96%) tenían tasas dentro de 2 desviaciones estándar. Generalmente, alrededor del 70% de los datos están dentro de 1 desviación estándar de la media y el 95% dentro de 2 desviaciones estándar, aunque esto está lejos de ser una regla estricta.

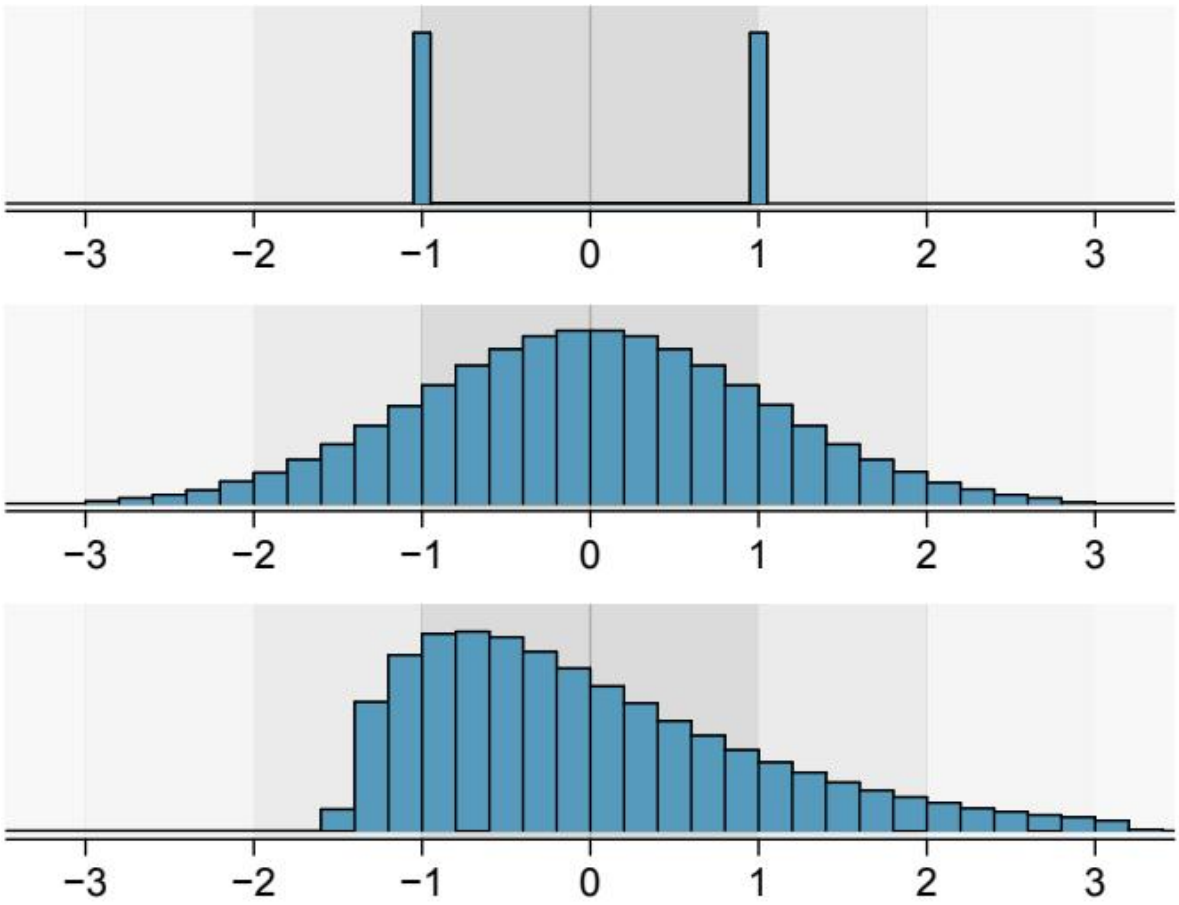


Figura 2.9: Tres distribuciones de población muy diferentes con la misma media $\mu = 0$ y desviación estándar $= 1$.

PRÁCTICA GUIADA 2.13

En la página 45, se introdujo el concepto de forma de una distribución. Una buena descripción de la forma de una distribución debería incluir la modalidad e indicar si la distribución es simétrica o sesgada hacia un lado. Usando la Figura 2.9 como ejemplo, explica por qué tal descripción es importante.9

EJEMPLO 2.14

Describe la distribución de la variable de tasa de interés utilizando el histograma en la Figura 2.6. La descripción debe incorporar el centro, la variabilidad y la forma de la distribución, y también debe ubicarse en contexto. También observa cualquier caso especialmente inusual.

La distribución de las tasas de interés es unimodal y está sesgada hacia el extremo superior. Muchas de las tasas se encuentran cerca de la media en 11.57%, y la mayoría se encuentra dentro de una desviación estándar (5.05%) de la media. Hay algunas tasas de interés excepcionalmente grandes en la muestra que están por encima del 20%.

En la práctica, la varianza y la desviación estándar a veces se utilizan como un medio para un fin, donde el "fin" es poder estimar con precisión la incertidumbre asociada con un estadístico de muestra. Por ejemplo, en el Capítulo 5 la desviación estándar se utiliza en cálculos que nos ayudan a comprender cuánto varía la media de una muestra de una muestra a otra.

9La Figura 2.9 muestra tres distribuciones que se ven bastante diferentes, pero todas tienen la misma media, varianza y desviación estándar. Usando la modalidad, podemos distinguir entre el primer gráfico (bimodal) y los dos últimos (unimodal). Usando la asimetría, podemos distinguir entre el último gráfico (sesgado a la derecha) y los dos primeros. Si bien una imagen, como un histograma, cuenta una historia más completa, podemos usar la modalidad y la forma (simetría/sesgo) para caracterizar información básica sobre una distribución.

2.1.5 Diagramas de caja, cuartiles y la mediana

Un diagrama de caja resume un conjunto de datos utilizando cinco estadísticas mientras que también grafica observaciones inusuales. La Figura 2.10 proporciona un diagrama de puntos vertical junto con un diagrama de caja de la variable de tasa de interés del conjunto de datos loan50.

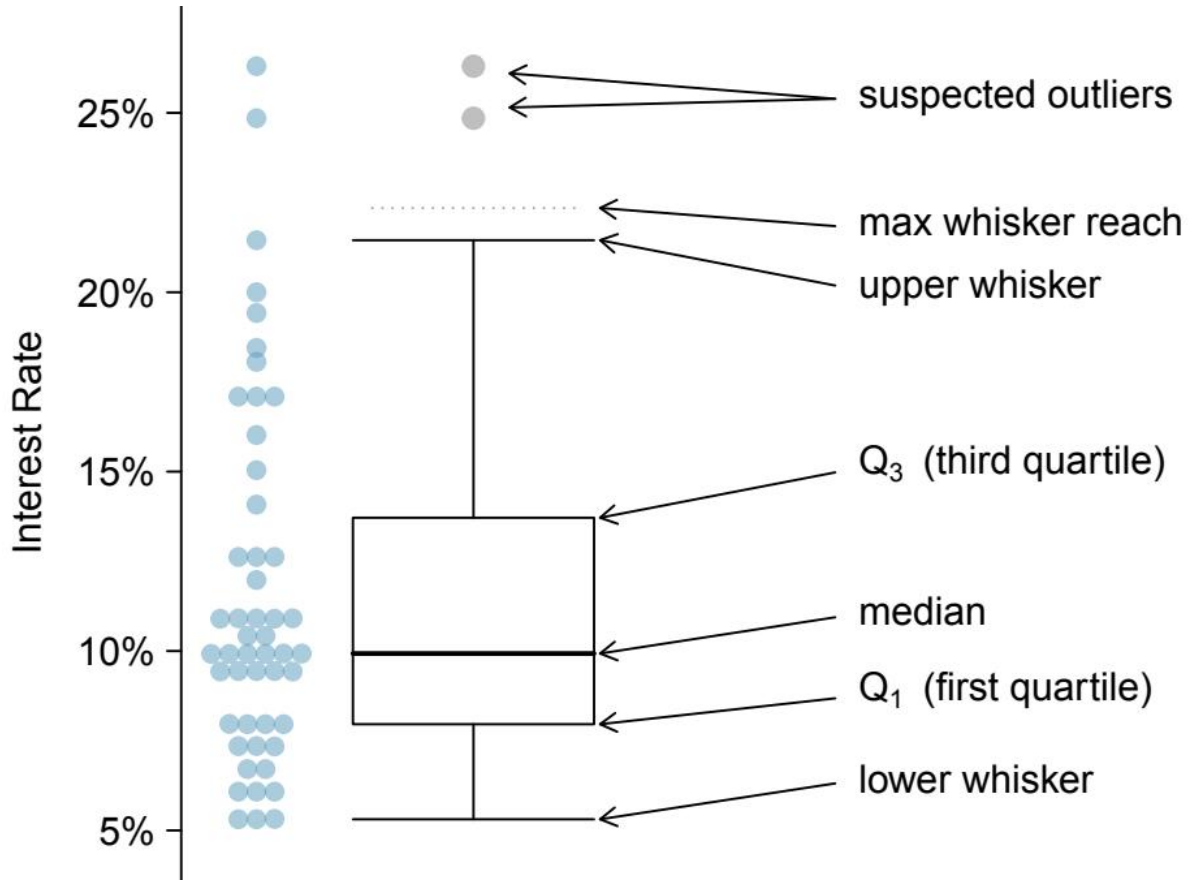


Figure 2.10: A vertical dot plot, where points have been horizontally stacked, next to a labeled box plot for the interest rates of the 50 loans.

El primer paso para construir un diagrama de caja es dibujar una línea oscura que denote la mediana, que divide los datos por la mitad. La Figura 2.10 muestra que el 50% de los datos cae por debajo de la mediana y el otro 50% cae por encima de la mediana. Hay 50 préstamos en el conjunto de datos (un número par), por lo que los datos se dividen perfectamente en dos grupos de 25. Tomamos la mediana en este caso como el promedio de las dos observaciones más cercanas al percentil 50, que resultan ser el mismo valor en este conjunto de datos: (9.93% + 9.93%)/2 = 9.93%. Cuando hay un número impar de observaciones, habrá exactamente una observación que divida los datos en dos mitades, y en tal caso esa observación es la mediana (no se necesita promedio).

MEDIANA: EL NÚMERO DE EN MEDIO

Si los datos están ordenados de menor a mayor, la mediana es la observación justo en el medio. Si hay un número par de observaciones, habrá dos valores en el medio, y la mediana se toma

como su promedio.

El segundo paso en la construcción de un diagrama de caja es dibujar un rectángulo para representar el 50% central de los datos. La longitud total de la caja, que se muestra verticalmente en la Figura 2.10, se llama rango intercuartílico (IQR, para abreviar). Éste, al igual que la desviación estándar, es una medida de la variabilidad de los datos. Cuanto más variables sean los datos, mayor tenderá a ser la desviación estándar y el IQR. Los dos límites de la caja se llaman primer cuartil (el percentil 25, es decir, el 25% de los datos cae por debajo de este valor) y el tercer cuartil (el percentil 75), y estos a menudo se etiquetan como Q1 y Q3, respectivamente.

RANGO INTERCUARTÍLICO (IQR)

El IQR es la longitud de la caja en un diagrama de caja. Se calcula como

$\text{IQR} = \text{Q3} - \text{Q1}$

donde Q1 y Q3 son los percentiles 25 y 75.

PRÁCTICA GUIADA 2.15

¿Qué porcentaje de los datos cae entre Q1 y la mediana? ¿Qué porcentaje está entre la mediana y Q3? 10

Extendiéndose desde la caja, los bigotes intentan capturar los datos fuera de la caja. Sin embargo, nunca se permite que su alcance sea superior a 1,5 × IQR. Capturan todo dentro de este alcance. En la Figura 2.10, el bigote superior no se extiende hasta los dos últimos puntos, que están más allá de Q3 + 1,5 × IQR, por lo que se extiende solo hasta el último punto por debajo de este límite. El bigote inferior se detiene en el valor más bajo, 5,31%, ya que no hay datos adicionales para alcanzar; el límite del bigote inferior no se muestra en la figura porque el gráfico no se extiende hasta Q1 −1,5×IQR. En cierto sentido, la caja es como el cuerpo del diagrama de caja y los bigotes son como sus brazos que intentan alcanzar el resto de los datos.

Cualquier observación que se encuentre más allá de los bigotes se etiqueta con un punto. El propósito de etiquetar estos puntos, en lugar de extender los bigotes a los valores mínimo y máximo observados, es ayudar a identificar cualquier observación que parezca inusualmente distante del resto de los datos. Las observaciones inusualmente distantes se denominan valores atípicos. En este caso, sería razonable clasificar las tasas de interés de 24,85% y 26,30% como valores atípicos, ya que están numéricamente distantes de la mayoría de los datos.

LOS VALORES ATÍPICOS SON EXTREMOS

Un valor atípico es una observación que parece extrema en relación con el resto de los datos.

El examen de los datos en busca de valores atípicos tiene muchos propósitos útiles, que incluyen

- 1. Identificar una fuerte asimetría en la distribución.
- 2. Identificar posibles errores en la recopilación o entrada de datos.
- 3. Proporcionar información sobre propiedades interesantes de los datos.

PRÁCTICA GUIADA 2.16

Usando la Figura 2.10, estime los siguientes valores para la tasa de interés en el conjunto de datos loan50: (a) Q1, (b) Q3 y (c) IQR.11

10Dado que Q1 y Q3 capturan el 50% central de los datos y la mediana divide los datos por la mitad, el 25% de los datos cae entre Q1 y la mediana, y otro 25% cae entre la mediana y Q3.

11Estas estimaciones visuales variarán un poco de una persona a otra: Q1 = 8%, Q3 = 14%, IQR = Q3 −Q1 = 6%. (Los valores verdaderos: Q1 = 7,96%, Q3 = 13,72%, IQR = 5,76%.)

2.1.6 Estadísticas robustas

¿Cómo se ven afectadas las estadísticas de muestra del conjunto de datos de la tasa de interés por la observación, 26,3%? ¿Qué habría pasado si este préstamo hubiera sido solo del 15%? ¿Qué pasaría con estas estadísticas resumidas si la observación del 26,3% hubiera sido aún mayor, digamos, del 35%? Estos escenarios se representan junto con los datos originales en la Figura 2.11, y las estadísticas de muestra se calculan en cada escenario en la Figura 2.12.

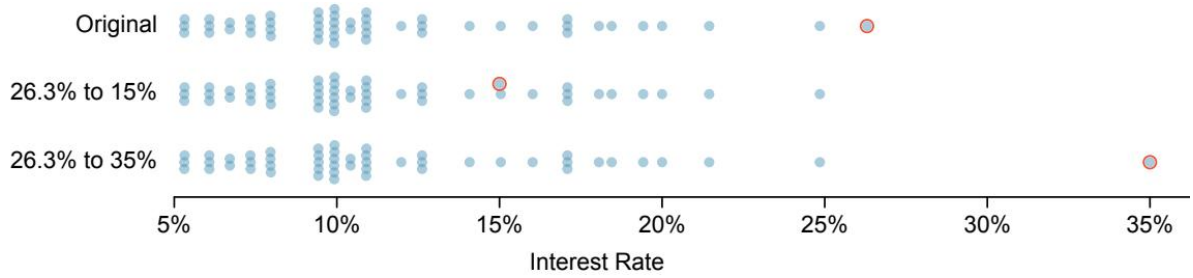


Figura 2.11: Gráficos de puntos de los datos originales de la tasa de interés y dos conjuntos de datos modificados.

	robusto		no robusto	
escenario	mediana	IQR	$\bar{x}$	s
datos originales de la tasa de interés	9.93%	5.76%	11.57%	5.05%
mover 26.3% → 15%	9.93%	5.76%	11.34%	4.61%
mover 26.3% → 35%	9.93%	5.76%	11.74%	5.68%

Figura 2.12: Una comparación de cómo la mediana, el IQR, la media ($\bar{x}$) y la desviación estándar (s) cambian si una observación extrema de la variable de la tasa de interés hubiera sido diferente.

PRÁCTICA GUIADA 2.17

(a) ¿Qué se ve más afectado por las observaciones extremas, la media o la mediana? La Figura 2.12 puede ser útil. (b) ¿La desviación estándar o el IQR se ven más afectados por las observaciones extremas?12

La mediana y el IQR se denominan estadísticas robustas porque las observaciones extremas tienen poco efecto sobre sus valores: mover el valor más extremo generalmente tiene poca influencia en estas estadísticas. Por otro lado, la media y la desviación estándar están más influenciadas por los cambios en las observaciones extremas, lo que puede ser importante en algunas situaciones.

EJEMPLO 2.18

La mediana y el IQR no cambiaron en los tres escenarios en la Figura 2.12. ¿Por qué podría ser este el caso?

La mediana y el IQR solo son sensibles a los números cercanos a Q1, la mediana y Q3. Dado que los valores en estas regiones son estables en los tres conjuntos de datos, las estimaciones de la mediana y el IQR también son estables.

PRÁCTICA GUIADA 2.19

La distribución de los montos de los préstamos en el conjunto de datos loan50 está sesgada a la derecha, con algunos préstamos grandes que permanecen en la cola derecha. Si quisiera comprender el tamaño típico del préstamo, ¿debería estar más interesado en la media o la mediana?13

12(a) La media se ve más afectada. (b) La desviación estándar se ve más afectada. Se proporcionan explicaciones completas en el material que sigue a la Práctica Guiada 2.17.

13¡Las respuestas variarán! Si simplemente buscamos comprender cómo es un préstamo individual típico, la mediana es probablemente más útil. Sin embargo, si el objetivo es comprender algo que se escala bien, como la cantidad total de dinero que podríamos necesitar tener a mano si ofreciéramos 1000 préstamos, entonces la media sería más útil.

2.1.7 Transformación de datos (tema especial)

Cuando los datos están muy fuertemente sesgados, a veces los transformamos para que sean más fáciles de modelar.

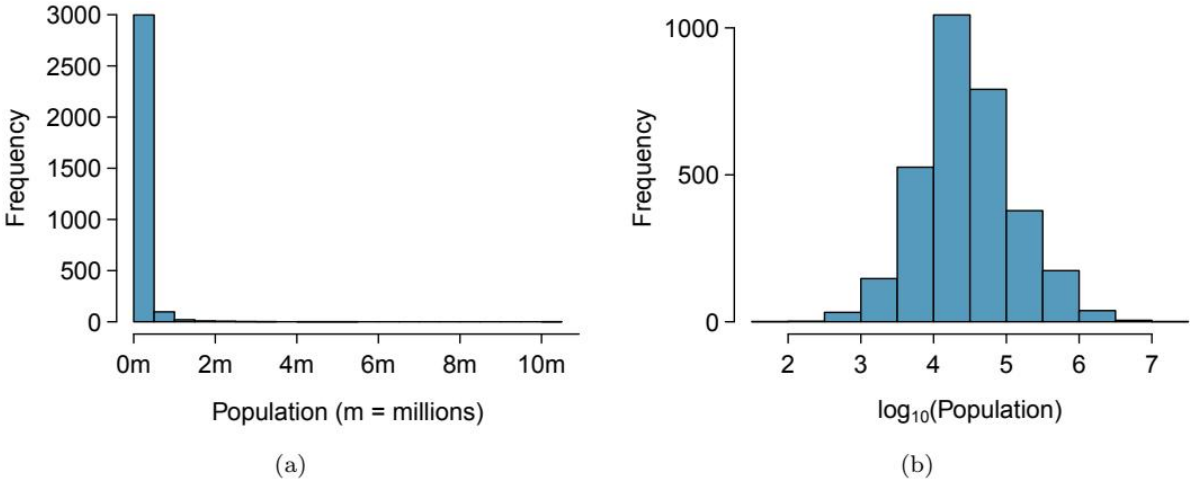


Figura 2.13: (a) Un histograma de las poblaciones de todos los condados de EE. UU. (b) Un histograma de las poblaciones de los condados transformadas con log10. Para este gráfico, el valor x corresponde a la potencia de 10, p. ej., "4" en el eje x corresponde a $104 = 10.000$.

EJEMPLO 2.20

Considere el histograma de las poblaciones de los condados que se muestra en la Figura 2.13(a), que muestra una asimetría extrema. ¿Qué no es útil de este gráfico?

Casi todos los datos caen en el bin más a la izquierda, y la asimetría extrema oscurece muchos de los detalles potencialmente interesantes de los datos.

Existen algunas transformaciones estándar que pueden ser útiles para datos fuertemente sesgados a la derecha donde gran parte de los datos son positivos pero están agrupados cerca de cero. Una transformación es un reescalado de los datos utilizando una función. Por ejemplo, un gráfico del logaritmo (base 10) de las poblaciones de los condados da como resultado el nuevo histograma en la Figura 2.13(b). Estos datos son simétricos y cualquier valor atípico potencial parece mucho menos extremo que en el conjunto de datos original. Al controlar los

valores atípicos y la asimetría extrema, las transformaciones como esta a menudo facilitan la construcción de modelos estadísticos contra los datos.

Las transformaciones también se pueden aplicar a una o ambas variables en un diagrama de dispersión. Un diagrama de dispersión del cambio de población de 2010 a 2017 frente a la población en 2010 se muestra en la Figura 2.14(a). En este primer diagrama de dispersión, es difícil descifrar patrones interesantes porque la variable de población está muy sesgada. Sin embargo, si aplicamos una transformación log10 a la variable de población, como se muestra en la Figura 2.14(b), se revela una asociación positiva entre las variables. De hecho, es posible que estemos interesados en ajustar una línea de tendencia a los datos cuando exploremos métodos para ajustar líneas de regresión en el Capítulo 8.

Otras transformaciones además del logaritmo también pueden ser útiles. Por ejemplo, la raíz cuadrada ($\sqrt{}$ observación original) y la inversa (1 observación original) son comúnmente utilizadas por los científicos de datos. Los objetivos comunes al transformar datos son ver la estructura de los datos de manera diferente, reducir la asimetría, ayudar en el modelado o enderezar una relación no lineal en un diagrama de dispersión.

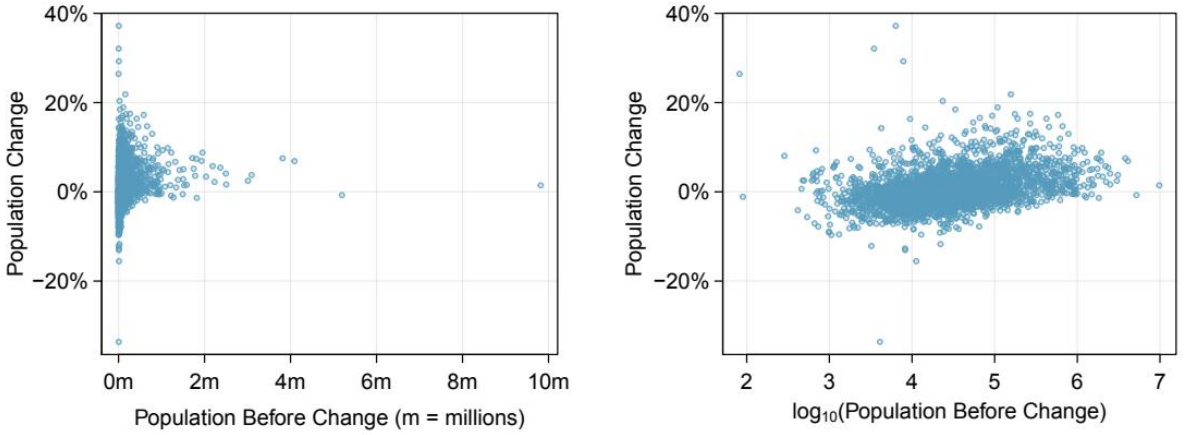


Figura 2.14: (a) Diagrama de dispersión del cambio de población frente a la población antes del cambio. (b) Un diagrama de dispersión de los mismos datos, pero donde el tamaño de la población se ha transformado logarítmicamente.

2.1.8 Mapeo de datos (tema especial)

El conjunto de datos del condado ofrece muchas variables numéricas que podríamos graficar usando diagramas de puntos, diagramas de dispersión o diagramas de caja, pero estos no capturan la verdadera naturaleza de los datos. Más bien, cuando encontramos datos geográficos, debemos crear un mapa de intensidad, donde los colores se utilizan para mostrar valores más altos y más bajos de una variable. Las figuras 2.15 y 2.16 muestran mapas de

intensidad para la tasa de pobreza en porcentaje (pobreza), la tasa de desempleo (tasa de desempleo), la tasa de propiedad de vivienda en porcentaje (propiedad de vivienda) y el ingreso medio por hogar (ingreso medio por hogar). La clave de color indica qué colores corresponden a qué valores. Los mapas de intensidad generalmente no son muy útiles para obtener valores precisos en un condado determinado, pero son muy útiles para ver tendencias geográficas y generar preguntas o hipótesis de investigación interesantes.

EJEMPLO 2.21

¿Qué características interesantes son evidentes en los mapas de intensidad de la pobreza y la tasa de desempleo?

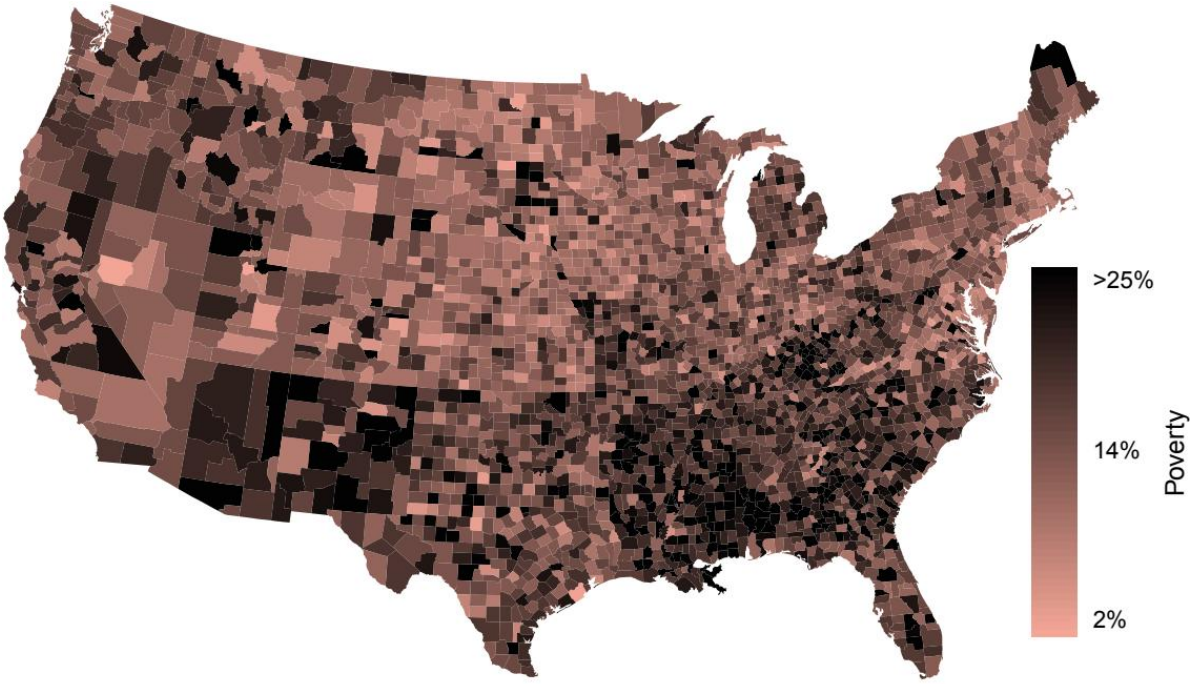
Las tasas de pobreza son evidentemente más altas en algunos lugares. En particular, el sur profundo muestra tasas de pobreza más altas, al igual que gran parte de Arizona y Nuevo México. Las altas tasas de pobreza son evidentes en las llanuras aluviales de Mississippi, un poco al norte de Nueva Orleans, y también en una gran sección de Kentucky.

La tasa de desempleo sigue tendencias similares, y podemos ver correspondencia entre las dos variables. De hecho, tiene sentido que las tasas más altas de desempleo estén estrechamente relacionadas con las tasas de pobreza. Una observación que destaca al comparar los dos mapas: la tasa de pobreza es mucho más alta que la tasa de desempleo, lo que significa que, aunque muchas personas pueden estar trabajando, no están ganando lo suficiente para salir de la pobreza.

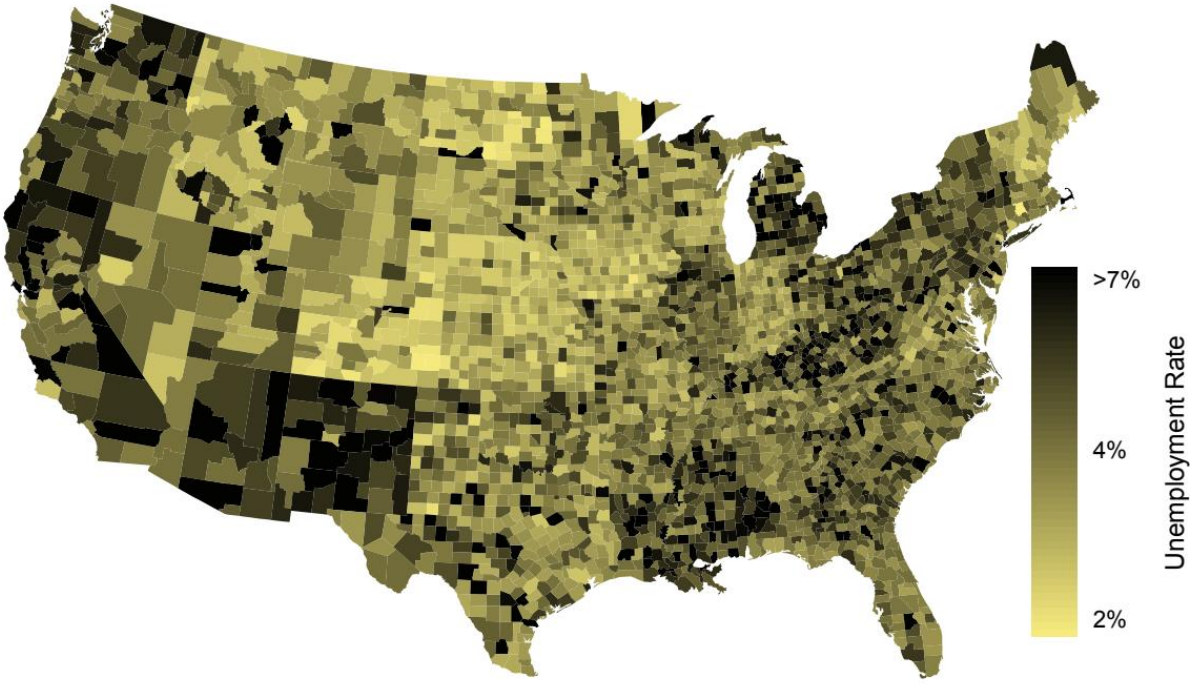
PRÁCTICA GUIADA 2.22

¿Qué características interesantes son evidentes en el mapa de intensidad del ingreso medio por hogar en la Figura 2.16(b)? 14

14Nota: las respuestas variarán. Existe cierta correspondencia entre las zonas de altos ingresos y las zonas metropolitanas, donde podemos ver puntos más oscuros (mayor ingreso medio por hogar), aunque hay varias excepciones. Puede buscar grandes ciudades con las que esté familiarizado e intentar localizarlas en el mapa como puntos oscuros.

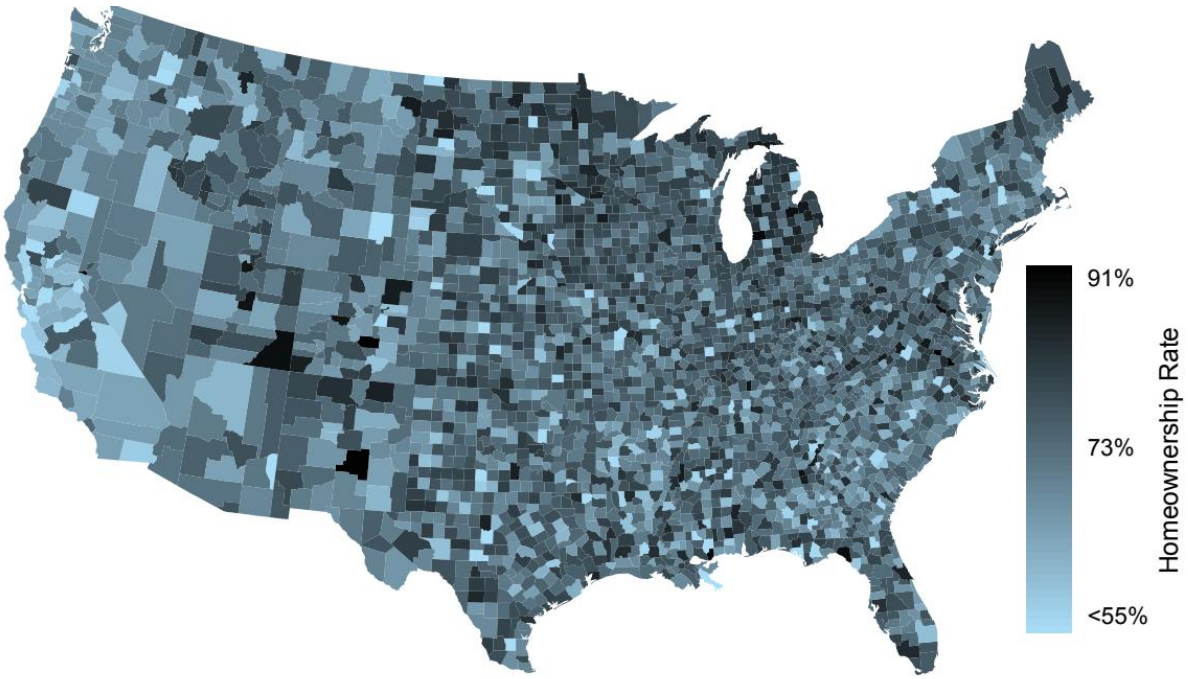


(a)

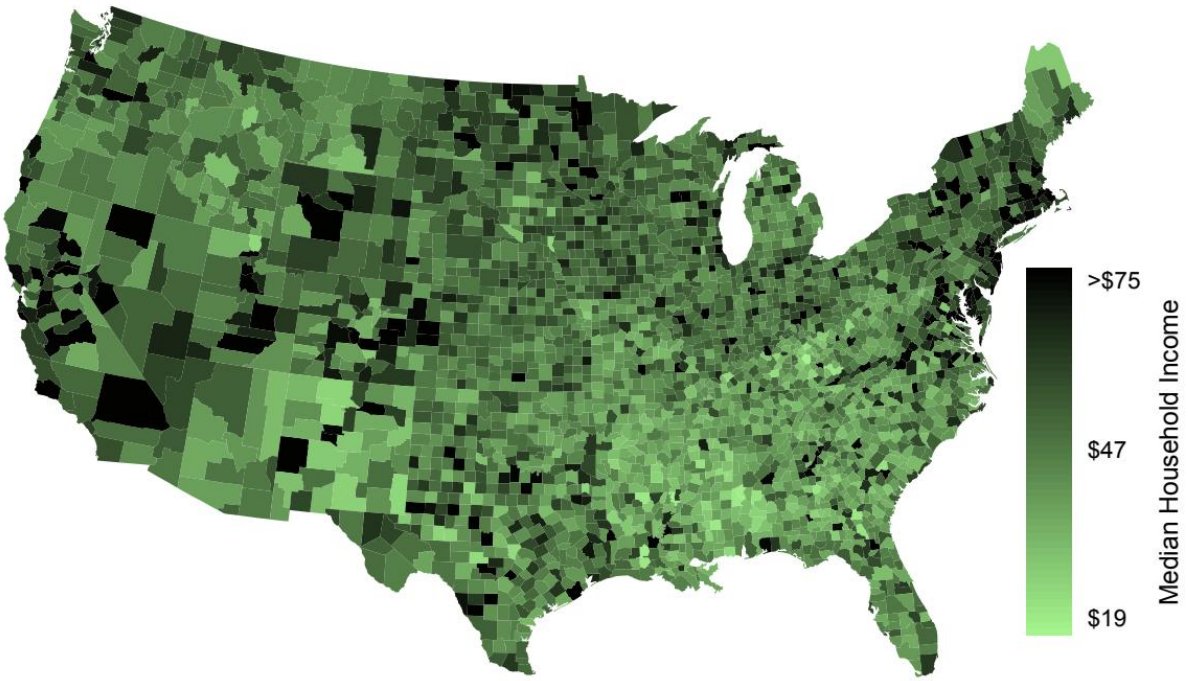


(b)

Figura 2.15: (a) Mapa de intensidad de la tasa de pobreza (porcentaje). (b) Mapa de la tasa de desempleo (porcentaje).



(a)



(b)

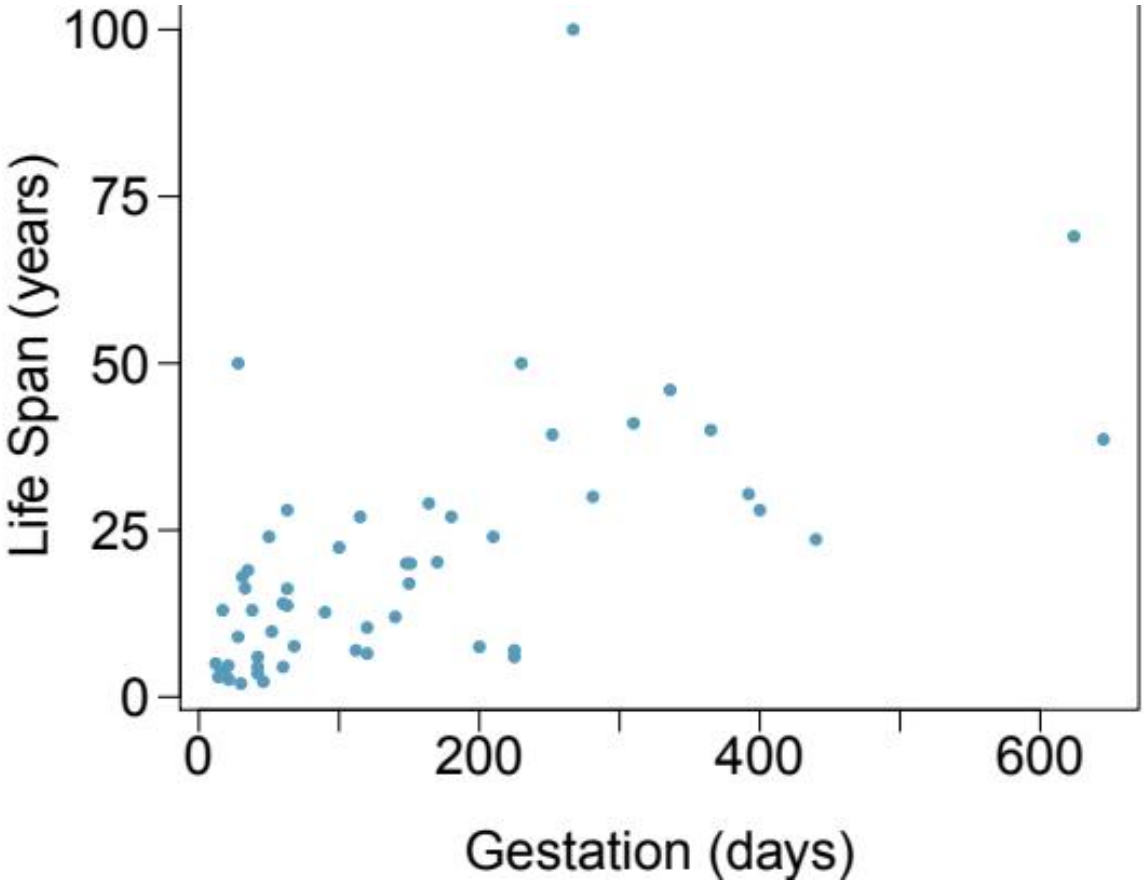
Figura 2.16: (a) Mapa de intensidad de la tasa de propiedad de vivienda (porcentaje). (b)

Mapa de intensidad del ingreso medio por hogar ($1000s).

Ejercicios

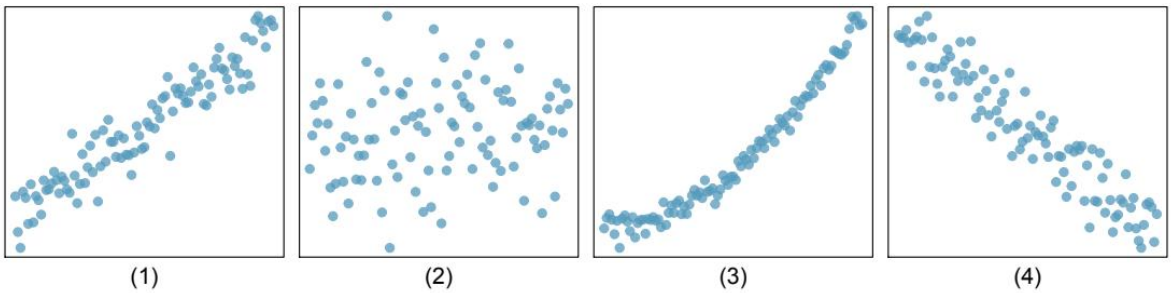
2.1 Esperanza de vida de mamíferos. Se recolectaron datos sobre la esperanza de vida (en años) y la duración de la gestación (en días) de 62 mamíferos. A continuación, se muestra un diagrama de dispersión de la esperanza de vida frente a la duración de la gestación.15

- (a) ¿Qué tipo de asociación es evidente entre la esperanza de vida y la duración de la gestación?
- (b) ¿Qué tipo de asociación esperaría ver si se invirtieran los ejes del gráfico, es decir, si graficáramos la duración de la gestación frente a la esperanza de vida?
- (c) ¿Son independientes la esperanza de vida y la duración de la gestación? Explique su razonamiento.



2.2 Asociaciones. Indique cuáles de los gráficos muestran (a) una asociación positiva, (b) una asociación negativa o (c) ninguna asociación. Determine también si las asociaciones positivas

y negativas son lineales o no lineales. Cada parte puede referirse a más de un gráfico.



2.3 Reproducción de bacterias. Suponga que solo hay suficiente espacio y nutrientes para mantener un millón de células bacterianas en una placa de Petri. Coloca algunas células bacterianas en esta placa de Petri, permite que se reproduzcan libremente y registra el número de células bacterianas en la placa a lo largo del tiempo. Dibuje un gráfico que represente la relación entre el número de células bacterianas y el tiempo.

2.4 Productividad de la oficina. La productividad de la oficina es relativamente baja cuando los empleados no sienten estrés por su trabajo o seguridad laboral. Sin embargo, los altos niveles de estrés también pueden conducir a una reducción de la productividad de los empleados. Dibuje un gráfico para representar la relación entre el estrés y la productividad.

2.5 Parámetros y estadísticas. Identifique qué valor representa la media muestral y qué valor representa la media poblacional reclamada.

- (a) Los hogares estadounidenses gastaron un promedio de aproximadamente $52 en 2007 en productos de Halloween, como disfraces, adornos y dulces. Para ver si este número había cambiado, los investigadores realizaron una nueva encuesta en 2008 antes de que se informaran los números de la industria. La encuesta incluyó a 1,500 hogares y encontró que el gasto promedio en Halloween fue de $58 por hogar.
- (b) El GPA promedio de los estudiantes en 2001 en una universidad privada fue de 3.37. Una encuesta en una muestra de 203 estudiantes de esta universidad arrojó un GPA promedio de 3.59 una década después.

2.6 Dormir en la universidad. Un artículo reciente en un periódico universitario decía que los estudiantes universitarios duermen un promedio de 5.5 horas cada noche. Un estudiante que se mostraba escéptico sobre este valor decidió realizar una encuesta muestreando aleatoriamente a 25 estudiantes. En promedio, los estudiantes muestreados durmieron 6.25 horas por noche. Identifique qué valor representa la media muestral y qué valor representa la media poblacional reclamada.

15T. Allison y D.V. Cicchetti. "Sleep in mammals: ecological and constitutional correlates". In: Arch. Hydrobiol 75 (1975), p. 442.

2.7 Días libres en una planta minera. Los trabajadores de un sitio minero en particular reciben un promedio de 35 días de vacaciones pagadas, lo cual es inferior al promedio nacional. El gerente de esta planta está bajo presión de un sindicato local para aumentar la cantidad de tiempo libre pagado. Sin embargo, no quiere dar más días libres a los trabajadores porque eso sería costoso. En cambio, decide que debería despedir a 10 empleados de tal manera que aumente el número promedio de días libres que informan sus empleados. Para lograr este objetivo, ¿debería despedir a los empleados que tienen la mayor cantidad de días libres, la menor cantidad de días libres o aquellos que tienen aproximadamente el número promedio de días libres?

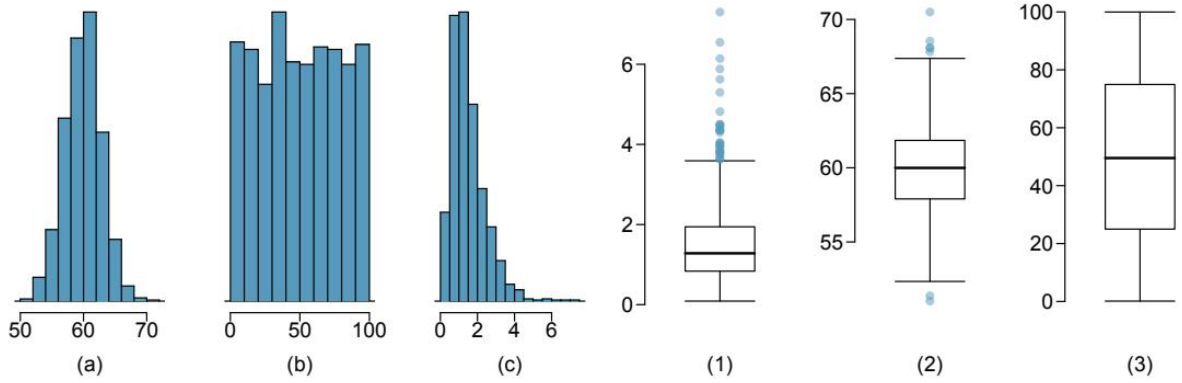
2.8 Medianas y RICs. Para cada parte, compare las distribuciones (1) y (2) según sus medianas y RICs. No es necesario que calcule estas estadísticas; simplemente indique cómo se comparan las medianas y los RICs. Asegúrese de explicar su razonamiento.

(a) (1) 3, 5, 6, 7, 9		(c) (1) 1, 2, 3, 4, 5
	(2) 3, 5, 6, 7, 20	(2) 6, 7, 8, 9, 10
(b) (1) 3, 5, 6, 7, 9		(d) (1) 0, 10, 50, 60, 100
	(2) 3, 5, 7, 8, 9	(2) 0, 100, 500, 600, 1000

2.9 Medias y DEs. Para cada parte, compare las distribuciones (1) y (2) según sus medias y desviaciones estándar. No es necesario que calcule estas estadísticas; simplemente indique cómo se comparan las medias y las desviaciones estándar. Asegúrese de explicar su razonamiento. Sugerencia: Puede ser útil dibujar diagramas de puntos de las distribuciones.

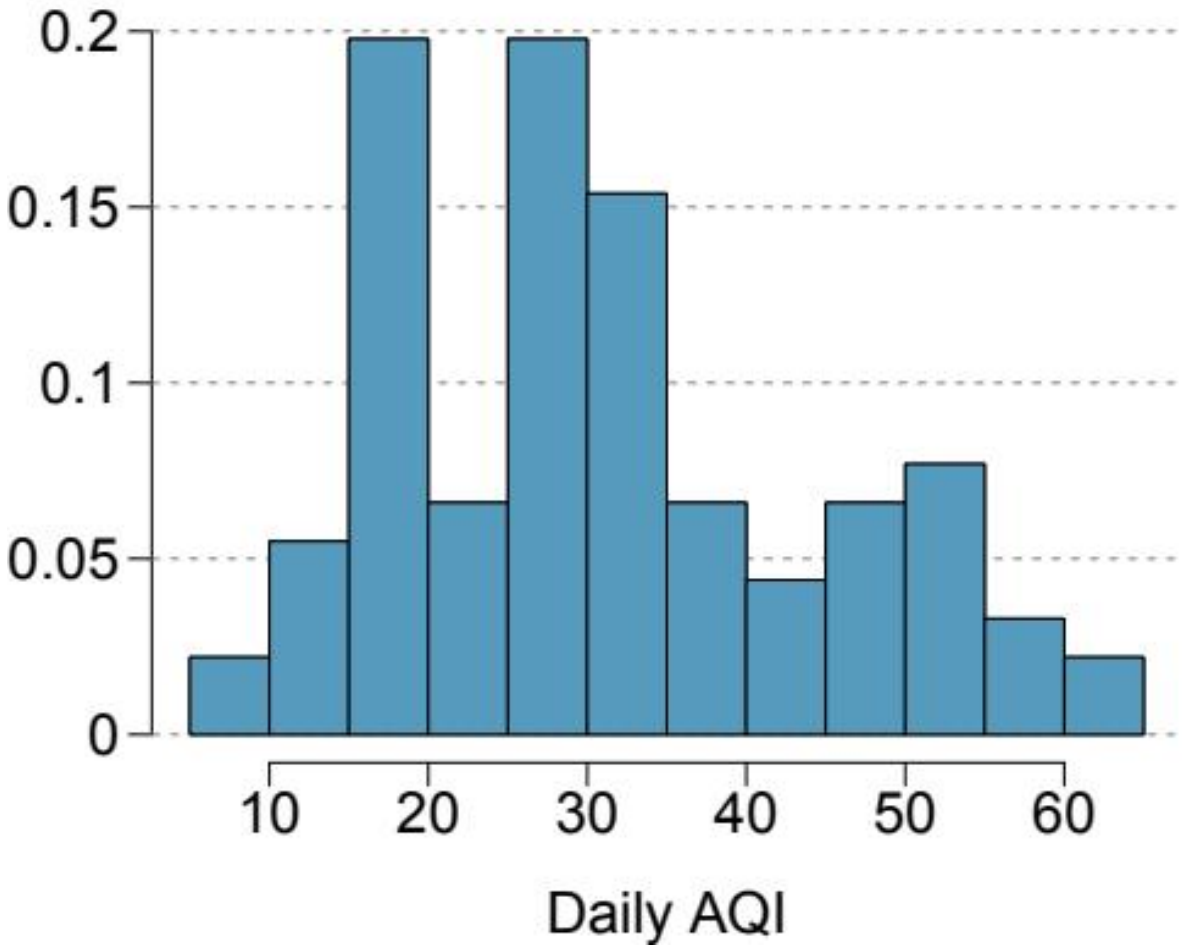
(a) (1) 3, 5, 5, 5, 8, 11, 11, 11, 13(2) 3, 5, 5, 5, 8, 11, 11, 11, 20	(c) (1) 0, 2, 4, 6, 8, 10(2) 20, 22, 24, 26, 28, 30
(b) (1) -20, 0, 0, 0, 15, 25, 30, 30(2) -40, 0, 0, 0, 15, 25, 30, 30	(d) (1) 100, 200, 300, 400, 500(2) 0, 50, 300, 550, 600

2.10 Mezclar y combinar. Describe la distribución en los histogramas a continuación y haz coincidir con los diagramas de caja.

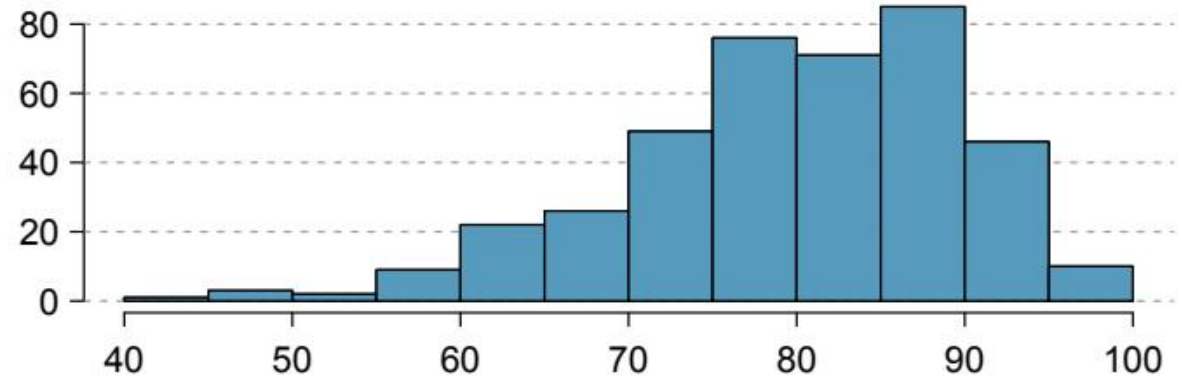


2.11 Calidad del aire. La calidad del aire diaria se mide mediante el índice de calidad del aire (ICA) informado por la Agencia de Protección Ambiental. Este índice informa el nivel de contaminación y qué efectos asociados en la salud podrían ser una preocupación. El índice se calcula para cinco contaminantes principales del aire regulados por la Ley de Aire Limpio y toma valores de 0 a 300, donde un valor más alto indica una menor calidad del aire. El ICA se informó para una muestra de 91 días en 2011 en Durham, NC.16

- (a) Estima el valor mediano del ICA de esta muestra.
- (b) ¿Esperaría que el valor medio del ICA de esta muestra sea mayor o menor que la mediana? Explica tu razonamiento.
- (c) Estima Q1, Q3 e IQR para la distribución.
- (d) ¿Alguno de los días de esta muestra se consideraría que tiene un ICA inusualmente bajo o alto? Explica tu razonamiento.

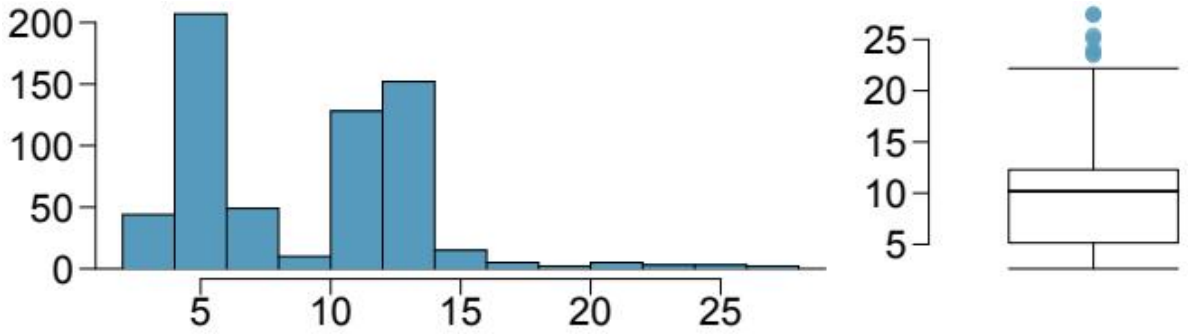


2.12 Mediana vs. media. Estima la mediana para las 400 observaciones que se muestran en el histograma y ten en cuenta si esperas que la media sea mayor o menor que la mediana.



2.13 Histogramas vs. diagramas de caja. Compare los dos diagramas a continuación.

¿Qué características de la distribución son evidentes en el histograma y no en el diagrama de caja? ¿Qué características son evidentes en el diagrama de caja pero no en el histograma?



2.14 Amigos de Facebook. Los datos de Facebook indican que el 50% de los usuarios de Facebook tienen 100 o más amigos, y que el número promedio de amigos de los usuarios es de 190. ¿Qué sugieren estos hallazgos sobre la forma de la distribución del número de amigos de los usuarios de Facebook?17

2.15 Distribuciones y estadísticas apropiadas, Parte I. Para cada uno de los siguientes, indica si esperas que la distribución sea simétrica, sesgada a la derecha o sesgada a la izquierda. Especifica también si la media o la mediana representarían mejor una observación típica en los datos, y si la variabilidad de las observaciones estaría mejor representada usando la desviación estándar o el IQR. Explica tu razonamiento.

- (a) Número de mascotas por hogar.
- (b) Distancia al trabajo, es decir, número de millas entre el trabajo y el hogar.
- (c) Alturas de hombres adultos.

16Agencia de Protección Ambiental de EE. UU., AirData, 2011.

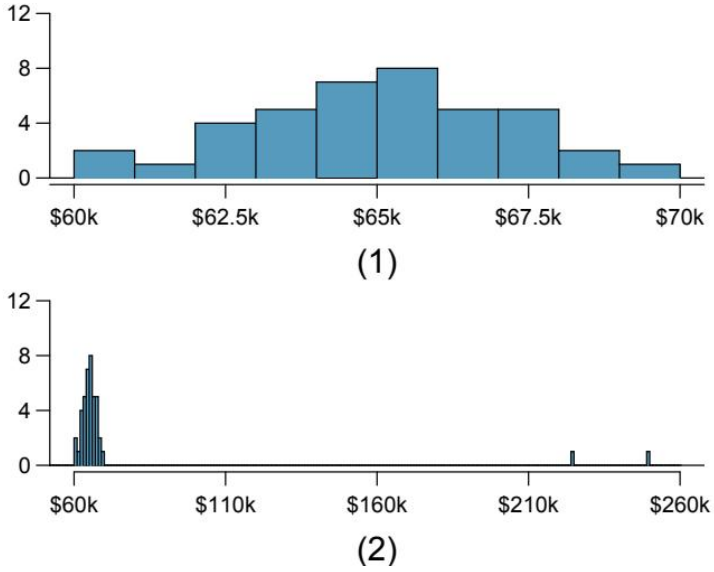
17Lars Backstrom. "Anatomía de Facebook". En: Notas del equipo de datos de Facebook (2011).

2.16 Distribuciones y estadísticas apropiadas, Parte II. Para cada uno de los siguientes, indica si esperas que la distribución sea simétrica, sesgada a la derecha o sesgada a la izquierda. Especifica también si la media o la mediana representarían mejor una observación típica en los datos, y si la variabilidad de las observaciones estaría mejor representada usando la desviación estándar o el IQR. Explica tu razonamiento.

- (a) Precios de la vivienda en un país donde el 25% de las casas cuestan menos de $350,000, el 50% de las casas cuestan menos de $450,000, el 75% de las casas cuestan menos de $1,000,000 y hay un número significativo de casas que cuestan más de $6,000,000.
- (b) Precios de la vivienda en un país donde el 25% de las casas cuestan menos de $300,000, el 50% de las casas cuestan menos de $600,000, el 75% de las casas cuestan menos de $900,000 y muy pocas casas cuestan más de $1,200,000.

- (c) Número de bebidas alcohólicas consumidas por estudiantes universitarios en una semana determinada. Suponga que la mayoría de estos estudiantes no beben porque son menores de 21 años, y solo unos pocos beben en exceso.
- (d) Salarios anuales de los empleados de una empresa Fortune 500 donde solo unos pocos ejecutivos de alto nivel ganan salarios mucho más altos que todos los demás empleados.

2.17 Ingresos en la cafetería. El primer histograma a continuación muestra la distribución de los ingresos anuales de 40 clientes en una cafetería universitaria. Suponga que dos nuevas personas entran en la cafetería: una que gana $225,000 y la otra $250,000. El segundo histograma muestra la nueva distribución de ingresos. También se proporcionan estadísticas resumidas.



	(1)	(2)
n	40	42
Min.	60,680	60,680
1st Qu.	63,620	63,710
Median	65,240	65,350
Mean	65,090	73,300
3rd Qu.	66,160	66,540
Max.	69,890	250,000
SD	2,122	37,321

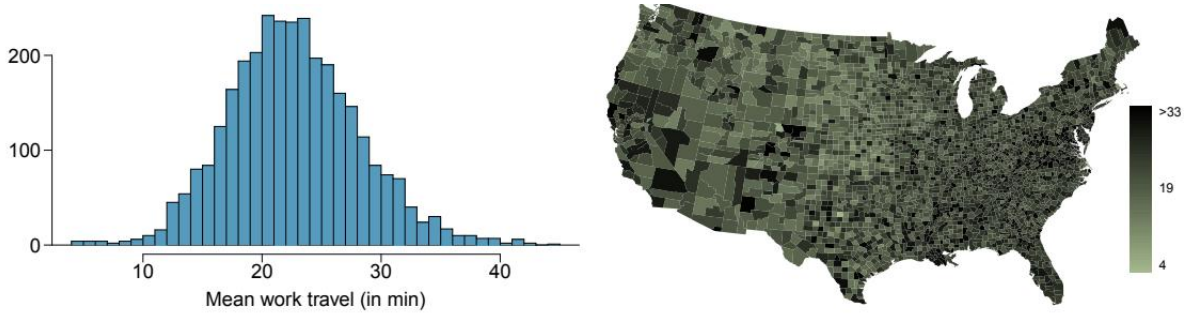
(a) ¿La media o la mediana representarían mejor lo que podríamos considerar un ingreso típico para los 42 clientes de esta cafetería? ¿Qué dice esto sobre la robustez de las dos medidas?

(b) ¿La desviación estándar o el IQR representarían mejor la cantidad de variabilidad en los ingresos de los 42 clientes de esta cafetería? ¿Qué dice esto sobre la robustez de las dos medidas?

2.18 Rango medio. El rango medio de una distribución se define como el promedio del máximo y el mínimo de esa distribución. ¿Es esta estadística robusta a los valores atípicos y al sesgo extremo? Explica tu razonamiento.

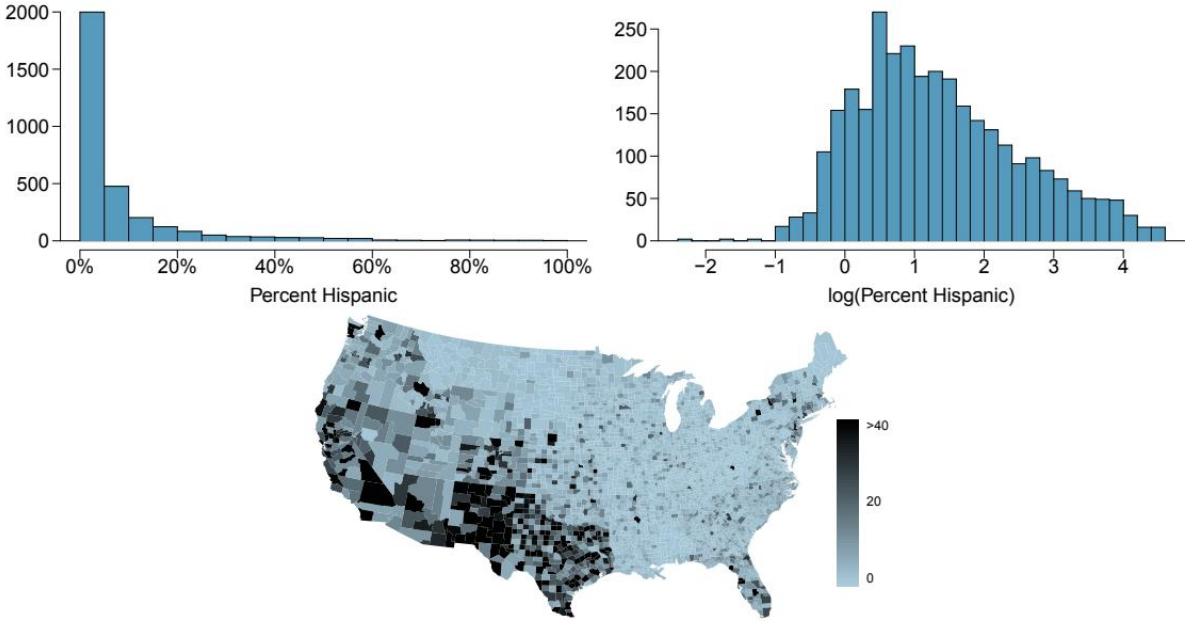
2.19 Tiempos de viaje. El censo de EE. UU. recopila datos sobre el tiempo que tardan los estadounidenses en viajar al trabajo, entre muchas otras variables. El histograma a continuación muestra la distribución de los tiempos promedio de viaje en 3,142 condados de

EE. UU. en 2010. También se muestra a continuación un mapa de intensidad espacial de los mismos datos.



- (a) Describe la distribución numérica y comenta si una transformación logarítmica podría ser aconsejable para estos datos.
- (b) Describe la distribución espacial de los tiempos de viaje utilizando el mapa anterior.

2.20 Población hispana. El censo de EE. UU. recopila datos sobre la raza y el origen étnico de los estadounidenses, entre muchas otras variables. El histograma a continuación muestra la distribución del porcentaje de la población que es hispana en 3,142 condados de los EE. UU. en 2010. También se muestra un histograma de los logaritmos de estos valores.



- (a) Describe la distribución numérica y comenta por qué podríamos querer usar valores transformados logarítmicamente al analizar o modelar estos datos.

- (b) ¿Qué características de la distribución de la población hispana en los condados de EE. UU. son evidentes en el mapa pero no en el histograma? ¿Qué características son evidentes en el histograma pero no en el mapa?
- (c) ¿Es una visualización más apropiada o útil que la otra? Explica tu razonamiento.

2.2 Considerando datos categóricos

En esta sección, presentaremos tablas y otras herramientas básicas para datos categóricos que se utilizan en todo este libro. El conjunto de datos loan50 representa una muestra de un conjunto de datos de préstamos más grande llamado loans. Este conjunto de datos más grande contiene información sobre 10,000 préstamos realizados a través de Lending Club. Examinaremos la relación entre la propiedad de la vivienda, que para los datos de los préstamos puede tomar un valor de alquiler, hipoteca (posee pero tiene una hipoteca) o propia, y el tipo de aplicación, que indica si la solicitud de préstamo se realizó con un socio o si era una solicitud individual.

2.2.1 Tablas de contingencia y diagramas de barras

La Figura 2.17 resume dos variables: tipo de solicitud y tenencia de vivienda. Una tabla que resume datos para dos variables categóricas de esta manera se llama tabla de contingencia. Cada valor en la tabla representa el número de veces que ocurrió una combinación particular de resultados de variables. Por ejemplo, el valor 3496 corresponde al número de préstamos en el conjunto de datos donde el prestatario alquila su vivienda y el tipo de solicitud fue individual. También se incluyen los totales de filas y columnas. Los totales de filas proporcionan los conteos totales en cada fila (p. ej., 3496 + 3839 + 1170 = 8505), y los totales de columnas son los conteos totales en cada columna. También podemos crear una tabla que muestre solo los porcentajes o proporciones generales para cada combinación de categorías, o podemos crear una tabla para una sola variable, como la que se muestra en la Figura 2.18 para la variable de tenencia de vivienda.

		Tenencia de Vivienda			
		alquiler	hipoteca	propia	Total
Tipo Sol.	individual	3496	3839	1170	8505
	conjunta	362	950	183	1495
	Total	3858	4789	1353	10000

Tenencia de Vivienda	Conteo
alquiler	3858

Tenencia de Vivienda	Conteo
hipoteca	4789
propia	1353
Total	10000

Figura 2.17: Una tabla de contingencia para el tipo de solicitud y la tenencia de vivienda.

Figura 2.18: Una tabla que resume las frecuencias de cada valor para la variable de tenencia de vivienda.

Un diagrama de barras es una forma común de mostrar una sola variable categórica. El panel izquierdo de la Figura 2.19 muestra un diagrama de barras para la variable de tenencia de vivienda. En el panel derecho, los conteos se convierten en proporciones, mostrando la proporción de observaciones que se encuentran en cada nivel (p. ej., 3858/10000 = 0.3858 para alquiler).

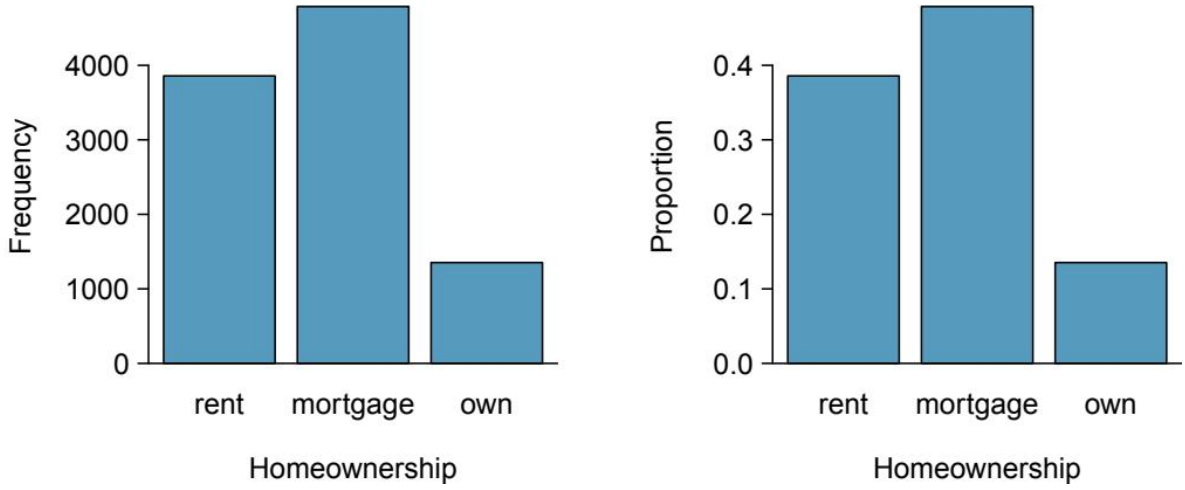


Figura 2.19: Dos diagramas de barras de número. El panel izquierdo muestra los conteos y el panel derecho muestra las proporciones en cada grupo.

2.2.2 Proporciones de fila y columna

A veces es útil comprender la descomposición fraccionaria de una variable en otra, y podemos modificar nuestra tabla de contingencia para proporcionar tal vista. La Figura 2.20 muestra las proporciones de fila para la Figura 2.17, que se calculan como los recuentos divididos por sus totales de fila. El valor 3496 en la intersección de individual y alquiler se reemplaza por 3496/8505 = 0.411, es decir, 3496 dividido por su total de fila, 8505. Entonces, ¿qué representa 0.411? Corresponde a la proporción de solicitantes individuales que alquilan.

	alquiler	hipoteca	propio	Total
individual	0.411	0.451	0.138	1.000
conjunta	0.242	0.635	0.122	1.000
Total	0.386	0.479	0.135	1.000

Figura 2.20: Una tabla de contingencia con proporciones de fila para las variables de tipo de solicitud y propiedad de la vivienda. El total de la fila está desviado en 0.001 para la fila conjunta debido a un error de redondeo.

Una tabla de contingencia de las proporciones de columna se calcula de manera similar, donde cada proporción de columna se calcula como el recuento dividido por el total de columna correspondiente. La Figura 2.21 muestra tal tabla, y aquí el valor 0.906 indica que el 90.6% de los inquilinos solicitaron el préstamo de forma individual. Esta tasa es más alta en comparación con los préstamos de personas con hipotecas (80.2%) o que son dueños de su casa (86.5%). Debido a que estas tasas varían entre los tres niveles de propiedad de la vivienda (alquiler, hipoteca, propio), esto proporciona evidencia de que las variables de tipo de solicitud y propiedad de la vivienda están asociadas.

	alquiler	hipoteca	propio	Total
individual	0.906	0.802	0.865	0.851
conjunta	0.094	0.198	0.135	0.150
Total	1.000	1.000	1.000	1.000

Figura 2.21: Una tabla de contingencia con proporciones de columna para las variables de tipo de solicitud y propiedad de la vivienda. El total de la última columna está desviado en 0.001 debido a un error de redondeo.

También podríamos haber verificado una asociación entre el tipo de solicitud y la propiedad de la vivienda en la Figura 2.20 utilizando las proporciones de fila. Al comparar estas proporciones de fila, miraríamos hacia abajo en las columnas para ver si la fracción de préstamos donde el prestatario alquila, tiene una hipoteca o es propietario variaba entre los tipos de solicitud individual y conjunta.

2.2. CONSIDERANDO DATOS CATEGÓRICOS 63

PRÁCTICA GUIADA 2.23

- (a) ¿Qué representa 0.451 en la Figura 2.20?
- (b) ¿Qué representa 0.802 en la Figura 2.21? 18

PRÁCTICA GUIADA 2.24

- (a) ¿Qué representa 0.122 en la intersección de "conjunto" y "propio" en la Figura 2.20?
- (b) ¿Qué representa 0.135 en la Figura 2.21? 19

EJEMPLO 2.25

Los científicos de datos utilizan la estadística para filtrar el spam de los mensajes de correo electrónico entrantes. Al observar características específicas de un correo electrónico, un científico de datos puede clasificar algunos correos electrónicos como spam o no spam con alta precisión. Una de estas características es si el correo electrónico no contiene números, números pequeños o números grandes. Otra característica es el formato del correo electrónico, que indica si un correo electrónico tiene o no contenido HTML, como texto en negrita. Nos centraremos en el formato del correo electrónico y el estado de spam utilizando el conjunto de datos de correo electrónico, y estas variables se resumen en una tabla de contingencia en la Figura 2.22. ¿Qué sería más útil para alguien que espera clasificar el correo electrónico como spam o correo electrónico normal para esta tabla: proporciones de fila o de columna?

Un científico de datos estaría interesado en cómo cambia la proporción de spam dentro de cada formato de correo electrónico. Esto corresponde a las proporciones de columna: la proporción de spam en los correos electrónicos de texto sin formato y la proporción de spam en los correos electrónicos HTML.

Si generamos las proporciones de columna, podemos ver que una mayor fracción de correos electrónicos de texto sin formato son spam (209/1195 = 17.5%) en comparación con los correos electrónicos HTML (158/2726 = 5.8%). Esta información por sí sola es insuficiente para clasificar un correo electrónico como spam o no spam, ya que más del 80% de los correos electrónicos de texto sin formato no son spam. Sin embargo, cuando combinamos cuidadosamente esta información con muchas otras características, tenemos una posibilidad razonable de poder clasificar algunos correos electrónicos como spam o no spam con confianza.

	texto	HTML	Total
spam	209	158	367
no spam	986	2568	3554
Total	1195	2726	3921

Figura 2.22: Una tabla de contingencia para spam y formato.

El ejemplo 2.25 señala que las proporciones de fila y columna no son equivalentes. Antes de decidirse por una forma para una tabla, es importante considerar cada una para asegurarse

de que se construya la tabla más útil. Sin embargo, a veces simplemente no está claro cuál, si alguna, es más útil.

EJEMPLO 2.26

Vuelva a mirar las Tablas 2.20 y 2.21. ¿Hay algún escenario obvio donde uno podría ser más útil que el otro?

¡Ninguno que pensáramos que fuera obvio! Lo que es distinto sobre el tipo de aplicación y la propiedad de la vivienda en comparación con el ejemplo del correo electrónico es que estas dos variables no tienen una clara relación de variable explicativa-respuesta que podríamos hipotetizar (vea la Sección 1.2.4 para estos términos). Por lo general, es más útil "condicionar" la variable explicativa. Por ejemplo, en el ejemplo del correo electrónico, el formato del correo electrónico se veía como una posible variable explicativa de si el mensaje era spam, por lo que nos resultaría más interesante calcular las frecuencias relativas (proporciones) para cada formato de correo electrónico.

18(a) 0.451 representa la proporción de solicitantes individuales que tienen una hipoteca. (b) 0.802 representa la fracción de solicitantes con hipotecas que solicitaron como individuos.

19(a) 0.122 representa la fracción de prestatarios conjuntos que son dueños de su casa. (b) 0.135 representa los prestatarios propietarios de viviendas que solicitaron conjuntamente el préstamo.

2.2.3 Uso de un diagrama de barras con dos variables

Las tablas de contingencia que utilizan proporciones de fila o columna son especialmente útiles para examinar cómo se relacionan dos variables categóricas. Los diagramas de barras apiladas proporcionan una forma de visualizar la información en estas tablas.

Un diagrama de barras apiladas es una visualización gráfica de la información de la tabla de contingencia. Por ejemplo, un diagrama de barras apiladas que representa la Figura 2.21 se muestra en la Figura 2.23(a), donde primero hemos creado un diagrama de barras utilizando la variable de propiedad de la vivienda y luego hemos dividido cada grupo por los niveles de tipo de aplicación.

Una visualización relacionada con el diagrama de barras apiladas es el diagrama de barras lado a lado, donde se muestra un ejemplo en la Figura 2.23(b).

Para el último tipo de diagrama de barras que presentamos, las proporciones de columna para la tabla de contingencia de tipo de aplicación y propiedad de vivienda se han traducido en un diagrama de barras apiladas estandarizado en la Figura 2.23(c). Este tipo de visualización es útil para comprender la fracción de solicitudes de préstamos individuales o conjuntas para prestatarios en cada nivel de propiedad de vivienda. Además, dado que las proporciones de

conjunto e individual varían entre los grupos, podemos concluir que las dos variables están asociadas.

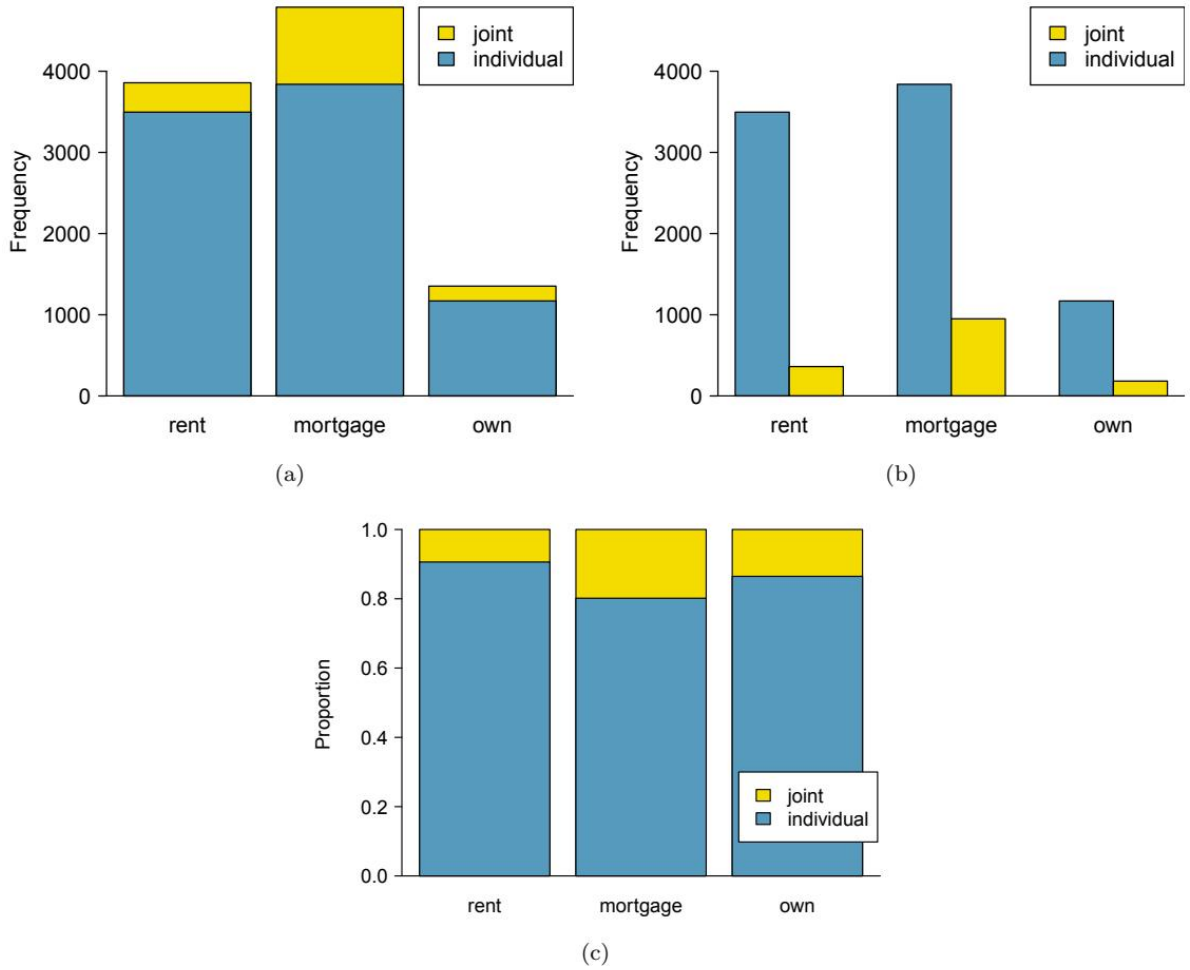


Figura 2.23: (a) Diagrama de barras apiladas para la propiedad de la vivienda, donde los conteos se han desglosado aún más por tipo de aplicación. (b) Diagrama de barras lado a lado para la propiedad de la vivienda y el tipo de aplicación. (c) Versión estandarizada del diagrama de barras apiladas.

2.2. CONSIDERANDO DATOS CATEGÓRICOS 65

EJEMPLO 2.27

Examine los tres gráficos de barras en la Figura 2.23. ¿Cuándo es más útil el gráfico de barras apiladas, lado a lado o apiladas estandarizadas?

El gráfico de barras apiladas es más útil cuando es razonable asignar una variable como variable explicativa y la otra variable como respuesta, ya que estamos agrupando efectivamente por una variable primero y luego dividiéndola por las otras.

Los gráficos de barras lado a lado son más agnósticos en su visualización sobre qué variable, si la hay, representa la variable explicativa y cuál la variable de respuesta. También es fácil discernir el número de casos en las seis combinaciones de grupos diferentes. Sin embargo, una desventaja es que tiende a requerir más espacio horizontal; la estrechez de la Figura 2.23(b) hace que el gráfico se sienta un poco apretado. Además, cuando dos grupos son de tamaños muy diferentes, como vemos en el grupo propio en relación con cualquiera de los otros dos grupos, es difícil discernir si existe una asociación entre las variables.

El gráfico de barras apiladas estandarizadas es útil si la variable primaria en el gráfico de barras apiladas está relativamente desequilibrada, por ejemplo, la categoría propia tiene solo un tercio de las observaciones en la categoría de hipoteca, lo que hace que el gráfico de barras apiladas simple sea menos útil para verificar una asociación. La principal desventaja de la versión estandarizada es que perdemos todo sentido de cuántos casos representa cada una de las barras.

2.2.4 Gráficos de Mosaico

Un gráfico de mosaico es una técnica de visualización adecuada para tablas de contingencia que se asemeja a un gráfico de barras apiladas estandarizado con el beneficio de que también vemos los tamaños de grupo relativos de la variable primaria.

Para comenzar a crear nuestro primer gráfico de mosaico, dividiremos un cuadrado en columnas para cada categoría de la variable de propiedad de la vivienda, con el resultado que se muestra en la Figura 2.24(a). Cada columna representa un nivel de propiedad de la vivienda, y los anchos de las columnas corresponden a la proporción de préstamos en cada una de esas categorías. Por ejemplo, hay menos préstamos donde el prestatario es propietario que donde el prestatario tiene una hipoteca. En general, los gráficos de mosaico utilizan áreas de caja para representar el número de casos en cada categoría.

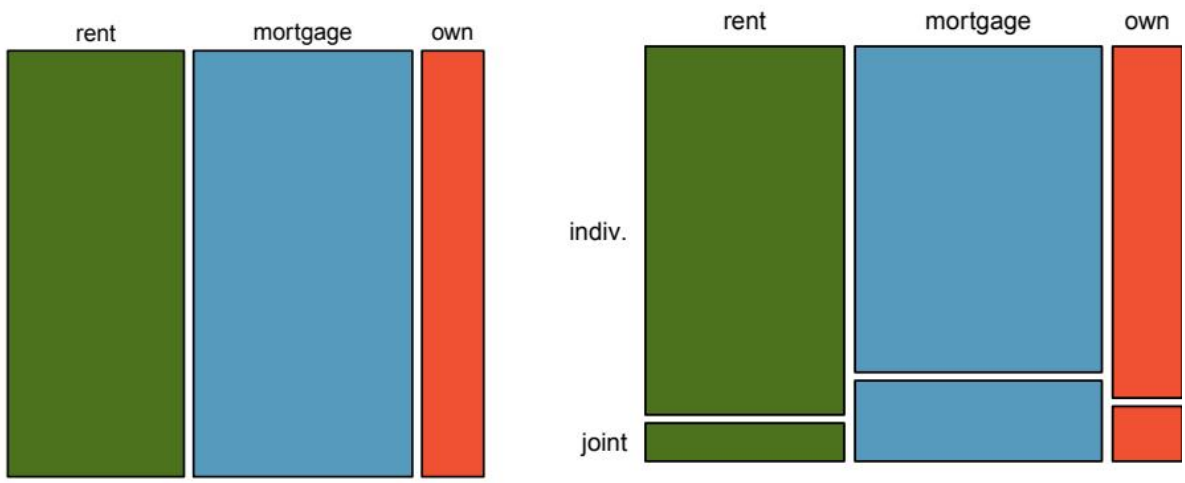


Figura 2.24: (a) El gráfico de mosaico de una variable para la propiedad de la vivienda. (b) Gráfico de mosaico de dos variables para la propiedad de la vivienda y el tipo de solicitud.

Para crear un gráfico de mosaico completo, el gráfico de mosaico de una sola variable se divide aún más en piezas en la Figura 2.24(b) utilizando la variable de tipo de solicitud. Cada columna se divide proporcionalmente al número de préstamos de prestatarios individuales y conjuntos. Por ejemplo, la segunda columna representa los préstamos en los que el prestatario tiene una hipoteca, y se dividió en préstamos individuales (superior) y préstamos conjuntos (inferior). Como otro ejemplo, el segmento inferior de la tercera columna representa los préstamos en los que el prestatario es dueño de su casa y solicitó conjuntamente, mientras que el segmento superior de esta columna representa a los prestatarios que son propietarios de viviendas y presentaron la solicitud individualmente. Podemos volver a utilizar este gráfico para ver que las variables de propiedad de la vivienda y tipo de solicitud están asociadas, ya que algunas columnas se dividen en diferentes

ubicaciones verticales que otras, que era la misma técnica utilizada para comprobar una asociación en el gráfico de barras apiladas estandarizado.

En la Figura 2.24, elegimos primero dividir por el estado de propietario de la vivienda del prestatario. Sin embargo, podríamos haber dividido primero por el tipo de solicitud, como en la Figura 2.25. Al igual que con los gráficos de barras, es común utilizar la variable explicativa para representar la primera división en un gráfico de mosaico, y luego que la respuesta divida cada nivel de la variable explicativa, si estas etiquetas son razonables para adjuntar a las variables en consideración.

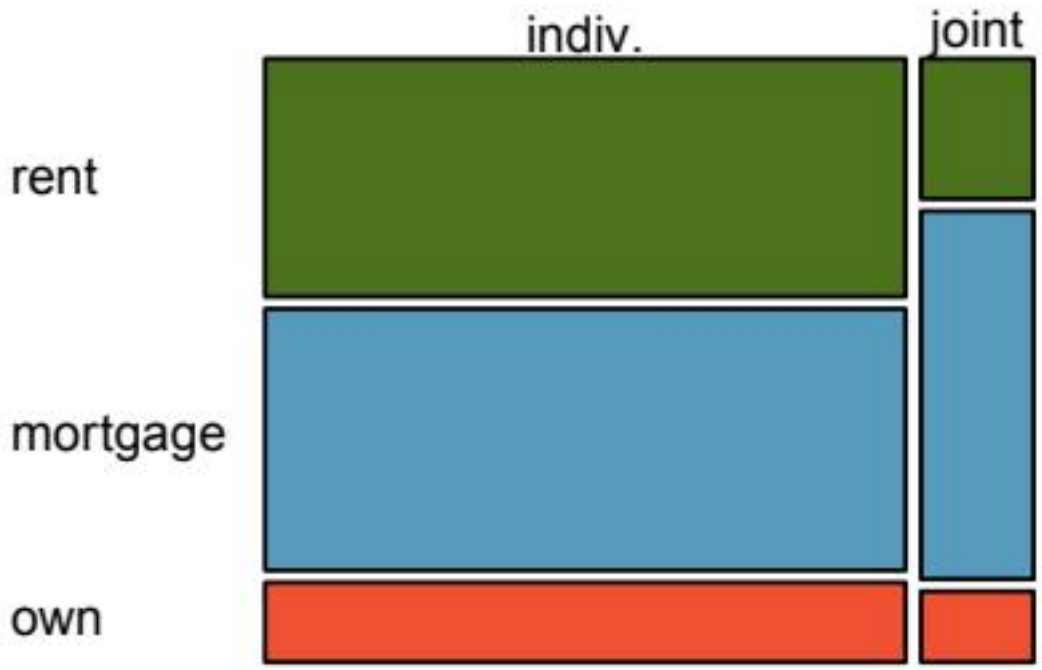


Figura 2.25: Gráfico de mosaico donde los préstamos se agrupan por la variable de propiedad de la vivienda después de haber sido divididos en los tipos de solicitud individual y conjunta.

2.2.5 El único gráfico circular que verás en este libro

Se muestra un gráfico circular en la Figura 2.26 junto con un diagrama de barras que representa la misma información. Los gráficos circulares pueden ser útiles para proporcionar una visión general de alto nivel que muestre cómo se desglosa un conjunto de casos. Sin embargo, también es difícil descifrar los detalles en un gráfico circular. Por ejemplo, se tarda un par de segundos más en reconocer que hay más préstamos donde el prestatario tiene una hipoteca que alquiler al mirar el gráfico circular, mientras que este detalle es muy obvio en el diagrama de barras. Si bien los gráficos circulares pueden ser útiles, preferimos los diagramas de barras por su facilidad para comparar grupos.

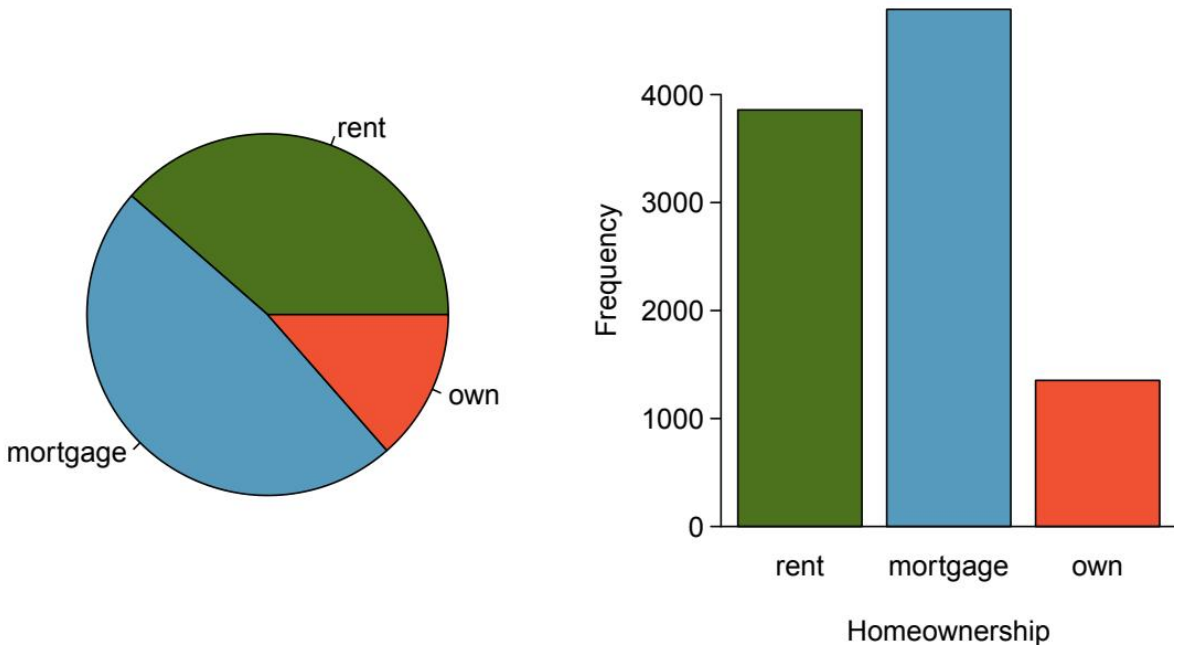


Figura 2.26: Un gráfico circular y un diagrama de barras de la propiedad de la vivienda.

2.2.6 Comparación de datos numéricos entre grupos

Algunas de las investigaciones más interesantes se pueden considerar examinando los datos numéricos entre grupos. Los métodos requeridos aquí no son realmente nuevos: todo lo que se requiere es hacer un gráfico numérico para cada grupo en el mismo gráfico. Aquí se introducen dos métodos convenientes: diagramas de caja lado a lado e histogramas huecos.

Volveremos a ver el conjunto de datos del condado y compararemos el ingreso familiar medio para los condados que ganaron población de 2010 a 2017 frente a los condados que no tuvieron ganancias. Si bien nos gustaría hacer una conexión causal aquí, recuerde que estos son datos de observación y, por lo tanto, tal interpretación sería, en el mejor de los casos, a medias.

Hubo 1,454 condados donde la población aumentó de 2010 a 2017, y hubo 1,672 condados sin ganancia (todos menos uno fueron una pérdida). Una muestra aleatoria de 100 condados del primer grupo y 50 del segundo grupo se muestra en la Figura 2.27 para dar una mejor idea de algunos de los datos brutos de ingresos medios.

Ganancia de población					Sin ganancia de población			
38.2	43.6	42.2	61.5	51.1	45.7	48.3	60.3	50.7
44.6	51.8	40.7	48.1	56.4	41.9	39.3	40.4	40.3
40.6	63.3	52.1	60.3	49.8	51.7	57	47.2	45.9
51.1	34.1	45.5	52.8	49.1	51	42.3	41.5	46.1
80.8	46.3	82.2	43.6	39.7	49.4	44.9	51.7	46.4
75.2	40.6	46.3	62.4	44.1	51.3	29.1	51.8	50.5
51.9	34.7	54	42.9	52.2	45.1	27	30.9	34.9
61	51.4	56.5	62	46	46.4	40.7	51.8	61.1
53.8	57.6	69.2	48.4	40.5	48.6	43.4	34.7	45.7
53.1	54.6	55	46.4	39.9	56.7	33.1	21	37
63	49.1	57.2	44.1	50	38.9	52	31.9	45.7
46.6	46.5	38.9	50.9	56	34.6	56.3	38.7	45.7
74.2	63	49.6	53.7	77.5	60	56.2	43	21.7
63.2	47.6	55.9	39.1	57.8	42.6	44.5	34.5	48.9
50.4	49	45.6	39	38.8	37.1	50.9	42.1	43.2
57.2	44.7	71.7	35.3	100.2		35.4	41.3	33.6
42.6	55.5	38.6	52.7	63		43.4	56.5	

Ingreso medio para 150 condados, en $1000

Figura 2.27: En esta tabla, se muestra el ingreso familiar medio (en $1000) de una muestra aleatoria de 100 condados que tuvieron ganancias de población a la izquierda. Los ingresos medios de una muestra aleatoria de 50 condados que no tuvieron ganancias de población se muestran a la derecha.

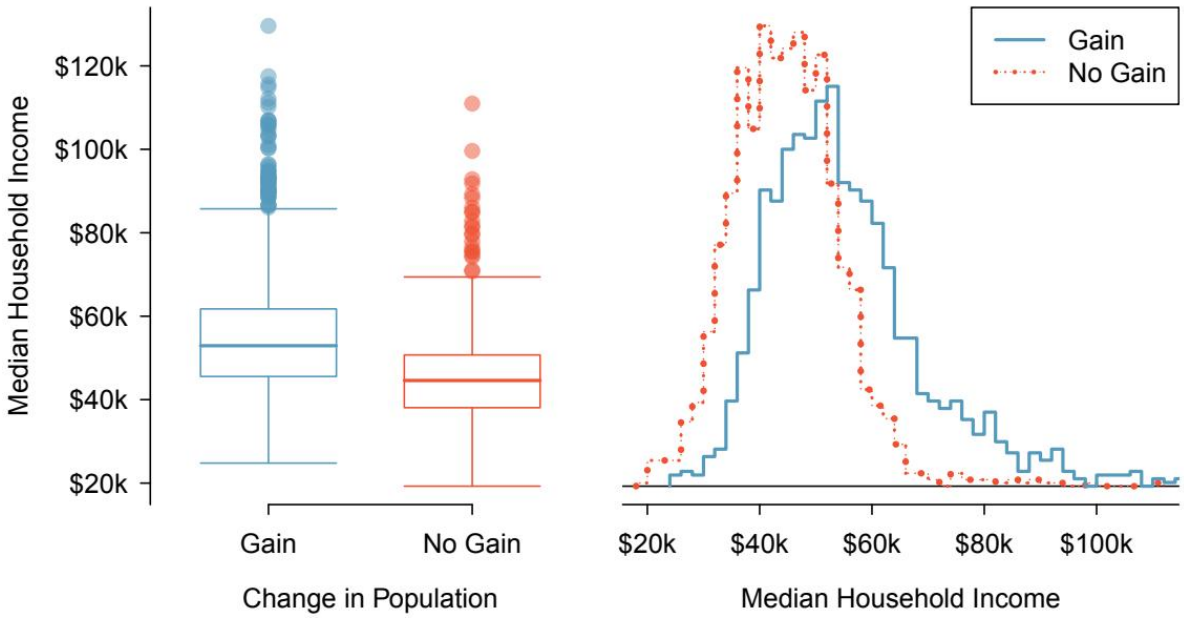


Figura 2.28: Diagrama de caja lado a lado (panel izquierdo) e histogramas huecos (panel derecho) para el ingreso familiar medio, donde los condados se dividen por si hubo una ganancia de población o si no hubo ganancia.

El diagrama de caja lado a lado es una herramienta tradicional para comparar entre grupos. Se muestra un ejemplo en el panel izquierdo de la Figura 2.28, donde hay dos diagramas de caja, uno para cada grupo, colocados en una ventana de trazado y dibujados en la misma escala.

Otro método de trazado útil utiliza histogramas huecos para comparar datos numéricos entre grupos. Estos son solo los contornos de los histogramas de cada grupo colocados en el mismo gráfico, como se muestra en el panel derecho de la Figura 2.28.

PRÁCTICA GUIADA 2.28

Utilice los gráficos de la Figura 2.28 para comparar los ingresos de los condados entre los dos grupos. ¿Qué nota sobre el centro aproximado de cada grupo? ¿Qué nota sobre la variabilidad entre grupos? ¿Es la forma relativamente consistente entre los grupos? ¿Cuántos modos prominentes hay para cada grupo?20

PRÁCTICA GUIADA 2.29

¿Qué componentes de cada gráfico en la Figura 2.28 le resultan más útiles?21

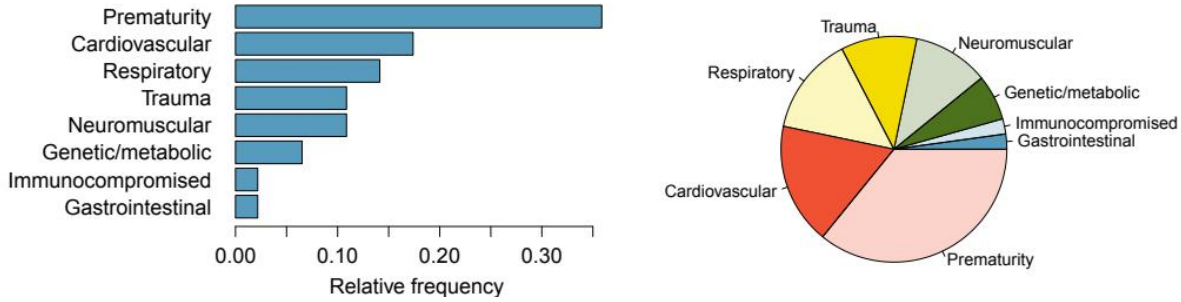
20Las respuestas pueden variar un poco. Los condados con ganancias de población tienden a tener ingresos más altos (mediana de alrededor de $45,000) en comparación con los condados

sin ganancias (mediana de alrededor de $40,000). La variabilidad también es ligeramente mayor para el grupo de ganancia de población. Esto es evidente en el IQR, que es aproximadamente un 50% mayor en el grupo de ganancia. Ambas distribuciones muestran una ligera a moderada asimetría hacia la derecha y son unimodales. Los diagramas de caja indican que hay muchas observaciones muy por encima de la mediana en cada grupo, aunque deberíamos anticipar que muchas observaciones caerán más allá de los bigotes al examinar cualquier conjunto de datos que contenga más de un par de cientos de puntos de datos.

21Las respuestas variarán. Los diagramas de caja lado a lado son especialmente útiles para comparar centros y extensiones, mientras que los histogramas huecos son más útiles para ver la forma de la distribución, la asimetría y las posibles anomalías.

Ejercicios

2.21 Uso de antibióticos en niños. El gráfico de barras y el gráfico circular a continuación muestran la distribución de las afecciones médicas preexistentes de los niños involucrados en un estudio sobre la duración óptima del uso de antibióticos en el tratamiento de la traqueítis, que es una infección del tracto respiratorio superior.



- (a) ¿Qué características son evidentes en el gráfico de barras pero no en el gráfico circular?
- (b) ¿Qué características son evidentes en el gráfico circular pero no en el gráfico de barras?
- (c) ¿Qué gráfico preferiría utilizar para mostrar estos datos categóricos?

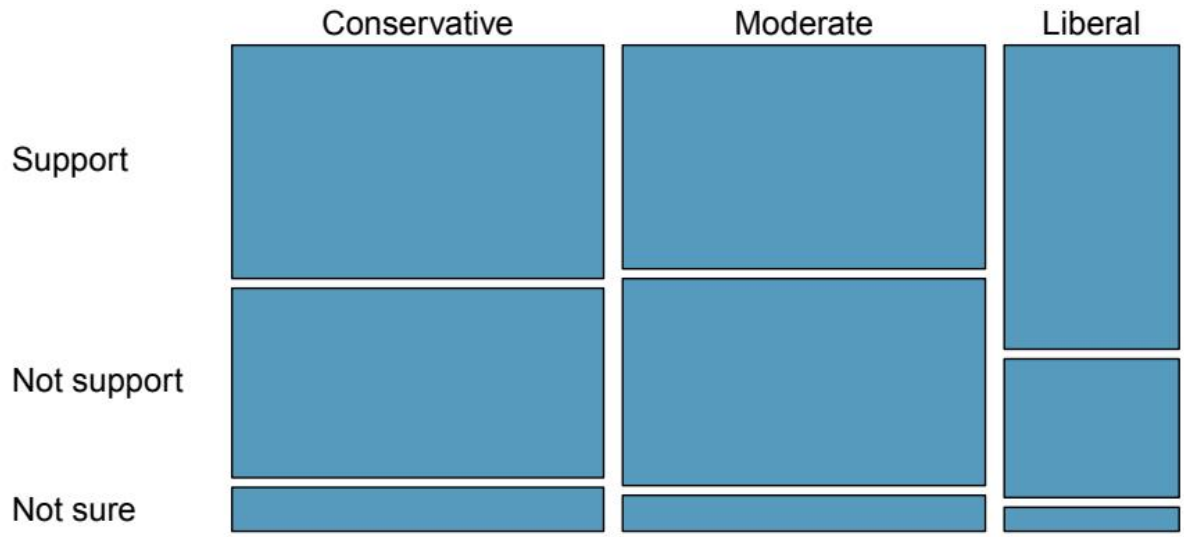
2.22 Opiniones sobre la inmigración. Se preguntó a 910 votantes registrados seleccionados al azar de Tampa, FL, si pensaban que los trabajadores que habían ingresado ilegalmente a los EE. UU. deberían (i) poder conservar sus trabajos y solicitar la ciudadanía estadounidense, (ii) poder conservar sus trabajos como trabajadores huéspedes temporales pero no poder solicitar la ciudadanía estadounidense, o (iii) perder sus trabajos y tener que salir del país. Los resultados de la encuesta por ideología política se muestran a continuación.22

		Ideología política			
		Conservador	Moderado	Liberal	Total
Respuesta	(i) Solicitar la ciudadanía	57	120	101	278
	(ii) Trabajador huésped	121	113	28	262
	(iii) Salir del país	179	126	45	350
	(iv) No estoy seguro	15	4	1	20
	Total	372	363	175	910

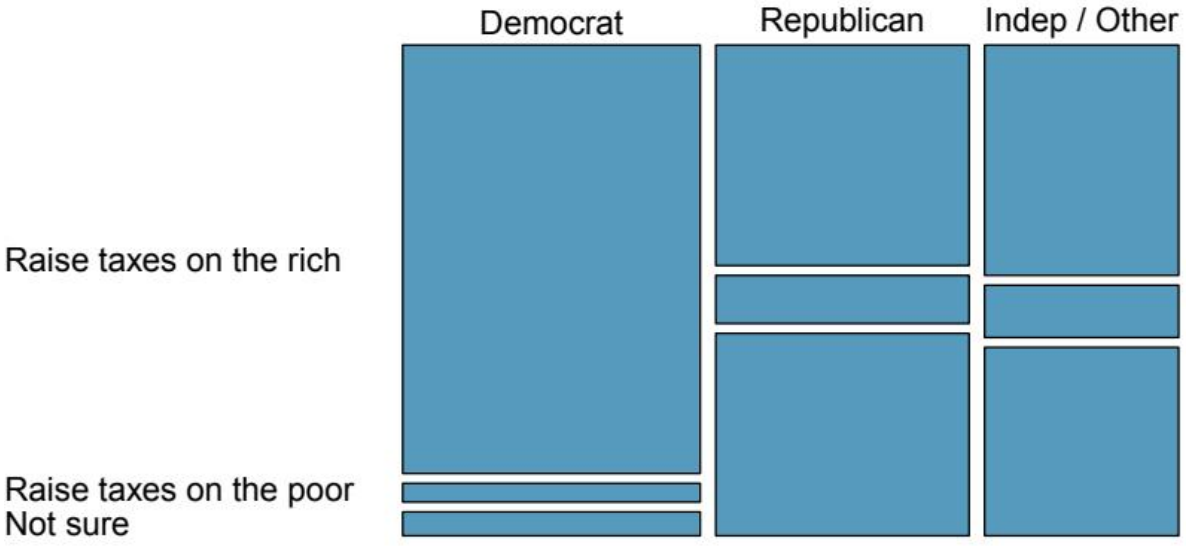
- (a) ¿Qué porcentaje de estos votantes de Tampa, FL, se identifican como conservadores?
- (b) ¿Qué porcentaje de estos votantes de Tampa, FL, está a favor de la opción de ciudadanía?
- (c) ¿Qué porcentaje de estos votantes de Tampa, FL, se identifican como conservadores y están a favor de la opción de ciudadanía?
- (d) ¿Qué porcentaje de estos votantes de Tampa, FL, que se identifican como conservadores también están a favor de la opción de ciudadanía? ¿Qué porcentaje de moderados comparte esta opinión? ¿Qué porcentaje de liberales comparte esta opinión?
- (e) ¿Parecen ser independientes la ideología política y las opiniones sobre la inmigración? Explique su razonamiento.

22SurveyUSA, Encuesta de noticias #18927, datos recopilados del 27 al 29 de enero de 2012.

2.23 Opiniones sobre la Ley DREAM. Se preguntó a una muestra aleatoria de votantes registrados de Tampa, FL, si apoyaban la Ley DREAM, una ley propuesta que proporcionaría un camino hacia la ciudadanía para las personas llevadas ilegalmente a los EE. UU. cuando eran niños. La encuesta también recopiló información sobre la ideología política de los encuestados. Según el diagrama de mosaico que se muestra a continuación, ¿parecen ser independientes las opiniones sobre la Ley DREAM y la ideología política? Explique su razonamiento.23



2.24 Aumento de impuestos. Se preguntó a una muestra aleatoria de votantes registrados a nivel nacional si pensaban que era mejor aumentar los impuestos a los ricos o aumentar los impuestos a los pobres. La encuesta también recopiló información sobre la afiliación al partido político de los encuestados. Según el diagrama de mosaico que se muestra a continuación, ¿parecen ser independientes las opiniones sobre el aumento de impuestos y la afiliación política? Explique su razonamiento.24



23SurveyUSA, Encuesta de noticias #18927, datos recopilados del 27 al 29 de enero de 2012.

24Public Policy Polling, Estadounidenses sobre títulos universitarios, literatura clásica, las estaciones y más, datos recopilados del 20 al 22 de febrero de 2015.

2.3 Estudio de caso: vacuna contra la malaria

EJEMPLO 2.30

Supongamos que su profesor divide a los estudiantes de la clase en dos grupos: los estudiantes de la izquierda y los estudiantes de la derecha. Si $\hat{}$pL y $\hat{}$pR representan la proporción de estudiantes que poseen un producto Apple a la izquierda y a la derecha, respectivamente, ¿le sorprendería que $\hat{}$pL no fuera exactamente igual a $\hat{}$pR ?

Si bien las proporciones probablemente serían cercanas entre sí, sería inusual que fueran exactamente iguales. Probablemente observaríamos una pequeña diferencia debido al azar.

PRÁCTICA GUIADA 2.31

Si no creemos que el lado de la sala en el que se sienta una persona en clase esté relacionado con si la persona posee o no un producto Apple, ¿qué suposición estamos haciendo sobre la relación entre estas dos variables?25

2.3.1 Variabilidad dentro de los datos

Consideramos un estudio sobre una nueva vacuna contra la malaria llamada PfSPZ. En este estudio, pacientes voluntarios fueron aleatorizados en uno de dos grupos experimentales: 14 pacientes recibieron una vacuna experimental y 6 pacientes recibieron una vacuna placebo. Diecinueve semanas después, los 20 pacientes fueron expuestos a una cepa de parásito de la malaria sensible a los fármacos; la motivación de usar una cepa de parásito sensible a los fármacos aquí es por consideraciones éticas, lo que permite que cualquier infección sea tratada eficazmente. Los resultados se resumen en la Figura 2.29, donde 9 de los 14 pacientes tratados permanecieron libres de signos de infección, mientras que los 6 pacientes del grupo de control mostraron algunos signos de infección basales.

		resultado		
	infecciónno infección			
tratamiento	vacuna	5	9	14
	placebo	6	0	6
	Total	11	9	20

Figura 2.29: Resultados resumidos para el experimento de la vacuna contra la malaria.

PRÁCTICA GUIADA 2.32

¿Es este un estudio observacional o un experimento? ¿Qué implicaciones tiene el tipo de estudio sobre lo que se puede inferir de los resultados?26

En este estudio, una proporción menor de pacientes que recibieron la vacuna mostraron signos de infección (35.7% versus 100%). Sin embargo, la muestra es muy pequeña y no está claro si la diferencia proporciona evidencia convincente de que la vacuna es eficaz.

25Estaríamos asumiendo que estas dos variables son independientes.

26El estudio es un experimento, ya que los pacientes fueron asignados aleatoriamente a un grupo experimental. Dado que este es un experimento, los resultados se pueden utilizar para evaluar una relación causal entre la vacuna contra la malaria y si los pacientes mostraron signos de infección.

EJEMPLO 2.33

A veces se solicita a los científicos de datos que evalúen la solidez de la evidencia. Al observar las tasas de infección de los pacientes en los dos grupos en este estudio, ¿qué se nos viene a la mente mientras intentamos determinar si los datos muestran evidencia convincente de una diferencia real?

Las tasas de infección observadas (35.7% para el grupo de tratamiento versus 100% para el grupo de control) sugieren que la vacuna puede ser eficaz. Sin embargo, no podemos estar seguros de si la diferencia observada representa la eficacia de la vacuna o si se debe simplemente al azar. Generalmente, hay un poco de fluctuación en los datos de la muestra, y no esperaríamos que las proporciones de la muestra sean exactamente iguales, incluso si la verdad fuera que las tasas de infección fueran independientes de la vacunación. Además, con muestras tan pequeñas, ¡quizás es común observar diferencias tan grandes cuando dividimos aleatoriamente un grupo debido solo al azar!

El ejemplo 2.33 es un recordatorio de que los resultados observados en la muestra de datos pueden no reflejar perfectamente las verdaderas relaciones entre las variables, ya que hay ruido aleatorio. Si bien la diferencia observada en las tasas de infección es grande, el tamaño de la muestra para el estudio es pequeño, lo que hace que no esté claro si esta diferencia observada representa la eficacia de la vacuna o si se debe simplemente al azar. Etiquetamos estas dos afirmaciones contrapuestas, H0 y HA, que se pronuncian como “H-cero” y “H-A”:

- H0: Modelo de independencia. Las variables tratamiento y resultado son independientes. No tienen relación, y la diferencia observada entre la proporción de pacientes que desarrollaron una infección en los dos grupos, 64.3%, se debió al azar.
- HA: Modelo alternativo. Las variables no son independientes. La diferencia en las tasas de infección del 64.3% no se debió al azar, y la vacuna afectó la tasa de infección.

¿Qué significaría si el modelo de independencia, que dice que la vacuna no tuvo influencia en la tasa de infección, fuera cierto? Significaría que 11 pacientes iban a desarrollar una infección sin importar en qué grupo fueran aleatorizados, y 9 pacientes no desarrollarían una infección sin importar en qué grupo fueran aleatorizados. Es decir, si la vacuna no afectara la tasa de infección, la diferencia en las tasas de infección se debería solo al azar en cómo se aleatorizaron los pacientes.

Ahora considere el modelo alternativo: las tasas de infección fueron influenciadas por si un paciente recibió la vacuna o no. Si esto fuera cierto, y especialmente si esta influencia fuera sustancial, esperaríamos ver alguna diferencia en las tasas de infección de los pacientes en los grupos.

Elegimos entre estas dos afirmaciones contrapuestas evaluando si los datos están tan en conflicto con H0 que el modelo de independencia no puede considerarse razonable. Si este es el caso, y los datos apoyan HA, entonces rechazaremos la noción de independencia y concluiremos que la vacuna fue eficaz.

2.3.2 Simulando el estudio

Vamos a implementar simulaciones, donde fingiremos que sabemos que la vacuna contra la malaria que se está probando no funciona. En última instancia, queremos entender si la gran diferencia que observamos es común en estas simulaciones. Si es común, entonces tal vez la diferencia que observamos se debió puramente al azar. Si es muy poco común, entonces la posibilidad de que la vacuna haya sido útil parece más plausible.

La Figura 2.29 muestra que 11 pacientes desarrollaron infecciones y 9 no. Para nuestra simulación, supondremos que las infecciones fueron independientes de la vacuna y que pudimos retroceder a cuando los investigadores aleatorizaron a los pacientes en el estudio. Si hubiéramos aleatorizado a los pacientes de manera diferente, podríamos obtener un resultado diferente en este mundo hipotético donde la vacuna no influye en la infección. Completemos otra aleatorización usando una simulación.

2.3. CASO DE ESTUDIO: VACUNA CONTRA LA MALARIA 73

En esta simulación, tomamos 20 fichas para representar a los 20 pacientes, donde escribimos "infección" en 11 fichas y "sin infección" en 9 fichas. En este mundo hipotético, creemos que cada paciente que contrajo una infección la iba a contraer independientemente del grupo en el que estuviera, así que veamos qué sucede si asignamos aleatoriamente a los pacientes a los grupos de tratamiento y control nuevamente. Barajamos a fondo las fichas y repartimos 14 en una pila de vacunas y 6 en una pila de placebo. Finalmente, tabulamos los resultados, que se muestran en la Figura 2.30.

		resultado		
		infección	no infección	Total
tratamiento	vacuna	7	7	14
(simulado)	placebo	4	2	6
	Total	11	9	20

Figura 2.30: Resultados de la simulación, donde cualquier diferencia en las tasas de infección se debe puramente al azar.

PRÁCTICA GUIADA 2.34

¿Cuál es la diferencia en las tasas de infección entre los dos grupos simulados en la Figura 2.30? ¿Cómo se compara esto con la diferencia observada del 64.3% en los datos reales?27

2.3.3 Comprobación de la independencia

En la Práctica Guiada 2.34, calculamos una posible diferencia bajo el modelo de independencia, que representa una diferencia debida al azar. Si bien en esta primera simulación, repartimos físicamente tarjetas para representar a los pacientes, es más eficiente realizar esta simulación usando una computadora. Al repetir la simulación en una computadora, obtenemos otra diferencia debida al azar:

$$\frac{2}{6} - \frac{9}{14} = -0.310$$

Y otra:

$$\frac{3}{6} - \frac{8}{14} = -0.071$$

Y así sucesivamente hasta que repitamos la simulación suficientes veces para tener una buena idea de lo que representa la distribución de las diferencias debidas únicamente al azar. La Figura 2.31 muestra un gráfico apilado de las diferencias encontradas a partir de 100 simulaciones, donde cada punto representa una diferencia simulada entre las tasas de infección (tasa de control menos tasa de tratamiento).

Observe que la distribución de estas diferencias simuladas está centrada alrededor de 0. Simulamos estas diferencias asumiendo que el modelo de independencia era verdadero, y bajo esta condición, esperamos que la diferencia esté cerca de cero con alguna fluctuación aleatoria,

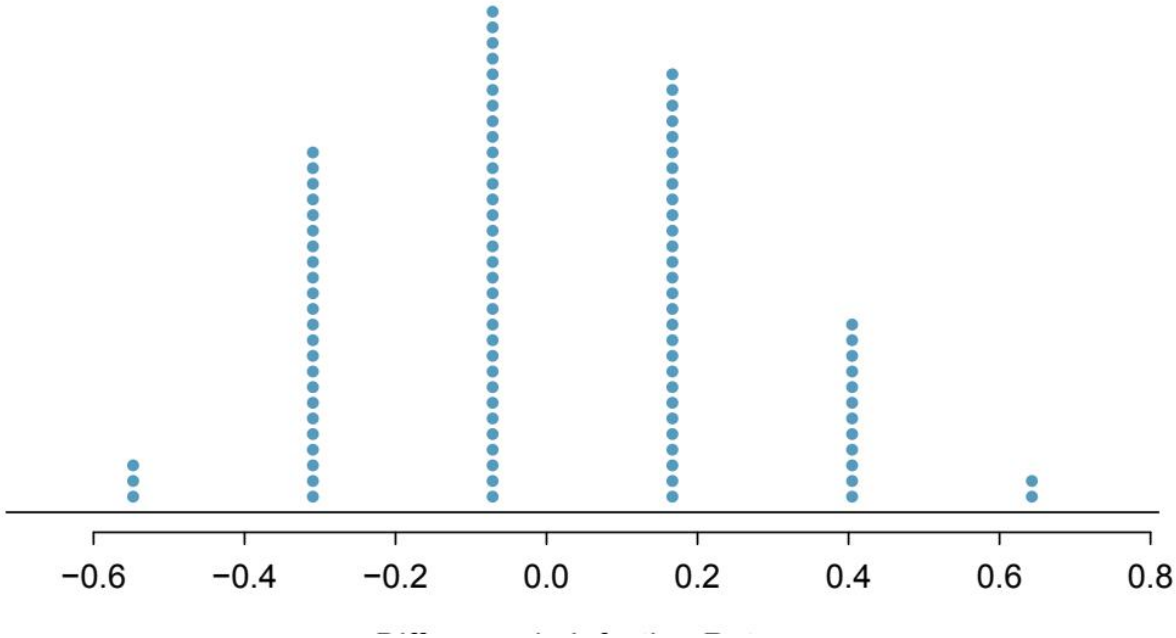
donde cerca es bastante generoso en este caso ya que los tamaños de muestra son tan pequeños en este estudio.

EJEMPLO 2.35

¿Con qué frecuencia observaría una diferencia de al menos 64.3% (0.643) según la Figura 2.31? ¿A menudo, a veces, raramente o nunca?

Parece que una diferencia de al menos 64.3% debida solo al azar ocurriría solo alrededor del 2% de las veces según la Figura 2.31. Una probabilidad tan baja indica un evento raro.

$274/6 - 7/14 = 0.167$ o aproximadamente 16.7% a favor de la vacuna. Esta diferencia debida al azar es mucho menor que la diferencia observada en los grupos reales.



Diferencia en las tasas de infección

Figura 2.31: Un diagrama de puntos apilados de las diferencias de 100 simulaciones producidas bajo el modelo de independencia, H0, donde en estas simulaciones las infecciones no se ven afectadas por la vacuna. Dos de las 100 simulaciones tuvieron una diferencia de al menos 64.3%, la diferencia observada en el estudio.

La diferencia del 64.3% como un evento raro sugiere dos posibles interpretaciones de los resultados del estudio:

- H0 Modelo de Independencia. La vacuna no tiene ningún efecto sobre la tasa de infección, y simplemente observamos una diferencia que solo ocurriría en una ocasión rara.

- HA Modelo Alternativo. La vacuna tiene un efecto sobre la tasa de infección, y la diferencia que observamos se debió en realidad a que la vacuna es eficaz para combatir la malaria, lo que explica la gran diferencia de 64.3%.

Basándonos en las simulaciones, tenemos dos opciones. (1) Concluimos que los resultados del estudio no proporcionan evidencia sólida en contra del modelo de independencia. Es decir, no tenemos evidencia suficientemente sólida para concluir que la vacuna tuvo un efecto en este entorno clínico. (2) Concluimos que la evidencia es suficientemente sólida para rechazar H0 y afirmar que la vacuna fue útil. Cuando realizamos estudios formales, generalmente rechazamos la noción de que simplemente observamos un evento raro.28 Entonces, en este caso, rechazamos el modelo de independencia en favor del alternativo. Es decir, estamos concluyendo que los datos proporcionan evidencia sólida de que la vacuna proporciona cierta protección contra la malaria en este entorno clínico.

Un campo de la estadística, la inferencia estadística, se basa en evaluar si tales diferencias se deben al azar. En la inferencia estadística, los científicos de datos evalúan qué modelo es más razonable dados los datos. Los errores ocurren, al igual que los eventos raros, y podríamos elegir el modelo equivocado. Si bien no siempre elegimos correctamente, la inferencia estadística nos brinda herramientas para controlar y evaluar con qué frecuencia ocurren estos errores. En el Capítulo 5, damos una introducción formal al problema de la selección de modelos. Pasamos los próximos dos capítulos construyendo una base de probabilidad y teoría necesaria para que esa discusión sea rigurosa.

28Este razonamiento no se extiende generalmente a las observaciones anecdóticas. Cada uno de nosotros observa eventos increíblemente raros todos los días, eventos que no podríamos esperar predecir. Sin embargo, en el entorno no riguroso de la evidencia anecdótica, casi cualquier cosa puede parecer un evento raro, por lo que la idea de buscar eventos raros en las actividades cotidianas es traicionera. Por ejemplo, podríamos mirar la lotería: ¡solo había 1 entre 292 millones de posibilidades de que los números de Powerball para el premio mayor más grande de la historia (13 de enero de 2016) fueran (04, 08, 19, 27, 34) con un Powerball de (10), pero sin embargo esos números salieron! Sin embargo, no importa qué números hubieran salido, habrían tenido las mismas probabilidades increíblemente raras. Es decir, cualquier conjunto de números que podríamos haber observado sería en última instancia increíblemente raro. Este tipo de situación es típico de nuestra vida diaria: cada evento posible en sí mismo parece increíblemente raro, pero si consideramos cada alternativa, esos resultados también son increíblemente raros. Debemos tener cuidado de no malinterpretar tal evidencia anecdótica.

Ejercicios

2.25 Efectos secundarios de Avandia. La rosiglitazona es el ingrediente activo de Avandia, un medicamento controvertido para la diabetes tipo 2, y se ha relacionado con un mayor riesgo de problemas cardiovasculares graves como accidentes cerebrovasculares, insuficiencia cardíaca

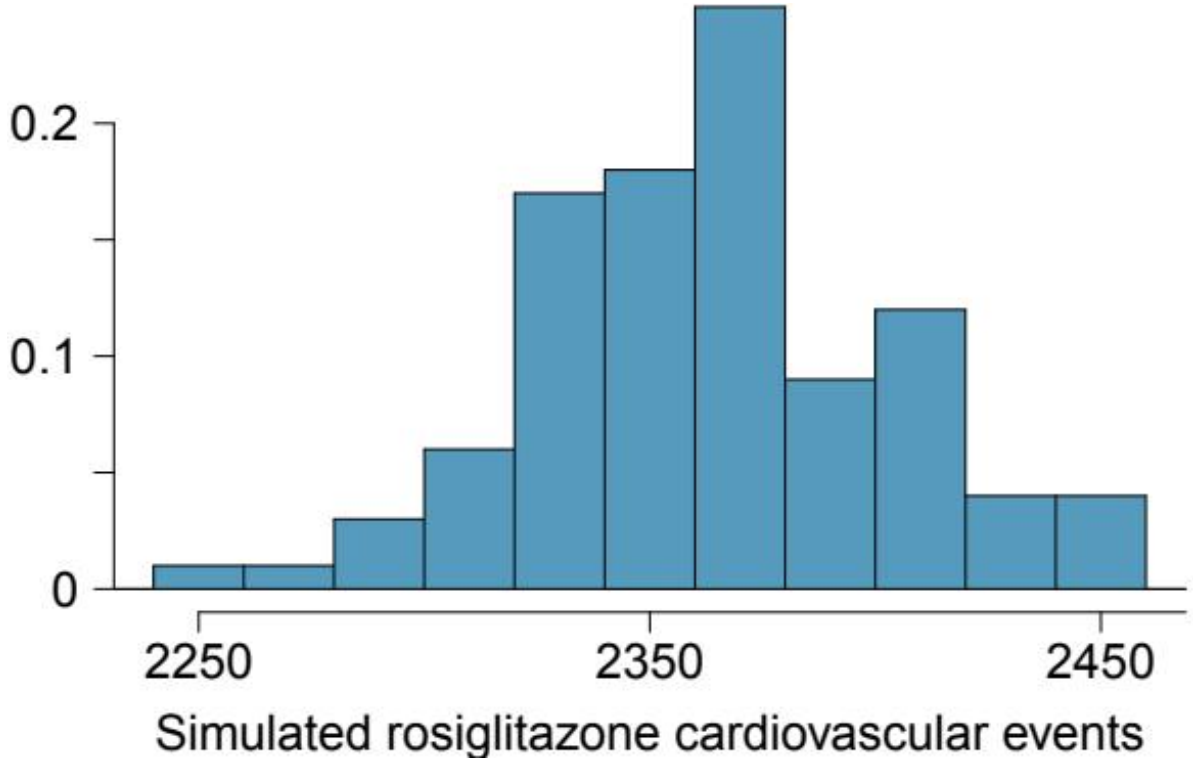
y muerte. Un tratamiento alternativo común es la pioglitazona, el ingrediente activo de un medicamento para la diabetes llamado Actos. En un estudio observacional retrospectivo a nivel nacional de 227,571 beneficiarios de Medicare de 65 años o más, se descubrió que 2,593 de los 67,593 pacientes que usaban rosiglitazona y 5,386 de los 159,978 que usaban pioglitazona tenían problemas cardiovasculares graves. Estos datos se resumen en la tabla de contingencia a continuación.29

		Problemas cardiovasculares		
		Sí	No	Total
Tratamiento	Rosiglitazona	2,593	65,000	67,593
	Pioglitazona	5,386	154,592	159,978
	Total	7,979	219,592	227,571

- (a) Determine si cada una de las siguientes afirmaciones es verdadera o falsa. Si es falsa, explique por qué. Tenga cuidado: el razonamiento puede ser incorrecto incluso si la conclusión de la afirmación es correcta. En tales casos, la afirmación debe considerarse falsa.
  - i. Dado que más pacientes con pioglitazona tuvieron problemas cardiovasculares (5,386 frente a 2,593), podemos concluir que la tasa de problemas cardiovasculares para aquellos en un tratamiento con pioglitazona es mayor.
  - ii. Los datos sugieren que los pacientes diabéticos que toman rosiglitazona tienen más probabilidades de tener problemas cardiovasculares, ya que la tasa de incidencia fue (2,593 / 67,593 = 0.038) 3.8% para los pacientes en este tratamiento, mientras que fue solo (5,386 / 159,978 = 0.034) 3.4% para los pacientes con pioglitazona.
  - iii. El hecho de que la tasa de incidencia sea mayor para el grupo de rosiglitazona prueba que la rosiglitazona causa problemas cardiovasculares graves.
  - iv. Según la información proporcionada hasta ahora, no podemos decir si la diferencia entre las tasas de incidencia se debe a una relación entre las dos variables o al azar.
- (b) ¿Qué proporción de todos los pacientes tuvo problemas cardiovasculares?
- (c) Si el tipo de tratamiento y tener problemas cardiovasculares fueran independientes, ¿aproximadamente cuántos pacientes en el grupo de rosiglitazona esperaríamos que hubieran tenido problemas cardiovasculares?
- (d) Podemos investigar la relación entre el resultado y el tratamiento en este estudio utilizando una técnica de aleatorización. Si bien en realidad realizaríamos las simulaciones necesarias para la aleatorización utilizando software estadístico, supongamos que en realidad simulamos utilizando tarjetas de índice. Para simular a partir del modelo de independencia, que establece que los resultados fueron independientes del tratamiento, escribimos si cada paciente tuvo o no un problema

cardiovascular en tarjetas, mezclamos todas las tarjetas y luego las repartimos en dos grupos de tamaño 67,593 y 159,978. Repetimos esta simulación 1,000 veces y cada vez registramos el número de personas en el grupo de rosiglitazona que tuvieron problemas cardiovasculares. Utilice el histograma de frecuencia relativa de estos conteos para responder (i)-(iii).

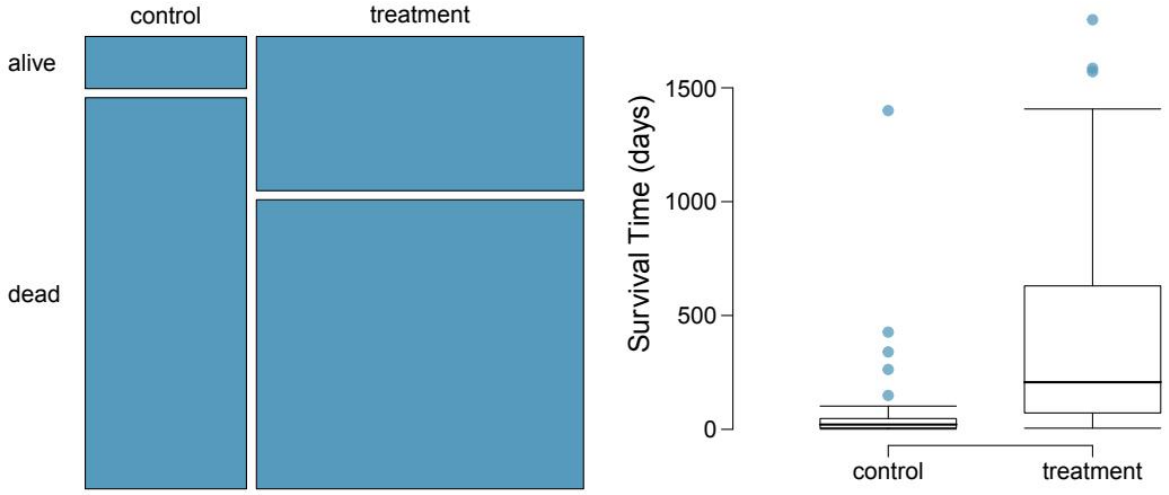
- i. ¿Cuáles son las afirmaciones que se están probando? ii. En comparación con el número calculado en la parte (c), ¿cuál proporcionaría más apoyo a la hipótesis alternativa, más o menos pacientes con problemas cardiovasculares en el grupo de rosiglitazona?
- iii. ¿Qué sugieren los resultados de la simulación sobre la relación entre tomar rosiglitazona y tener problemas cardiovasculares en pacientes diabéticos?



29D.J. Graham et al. “Risk of acute myocardial infarction, stroke, heart failure, and death in elderly Medicare patients treated with rosiglitazone or pioglitazone”. In: JAMA 304.4 (2010), p. 411. issn: 0098-7484.

2.26 Trasplantes de corazón. El Estudio de Trasplante de Corazón de la Universidad de Stanford se llevó a cabo para determinar si un programa experimental de trasplante de corazón aumentaba la esperanza de vida. Cada paciente que ingresaba al programa era designado candidato oficial para trasplante de corazón, lo que significaba que estaba gravemente enfermo y que lo más probable es que se beneficiara de un corazón nuevo. Algunos pacientes recibieron

un trasplante y otros no. La variable trasplante indica en qué grupo estaban los pacientes; los pacientes del grupo de tratamiento recibieron un trasplante y los del grupo de control no. De los 34 pacientes en el grupo de control, 30 murieron. De las 69 personas en el grupo de tratamiento, 45 murieron. Se utilizó otra variable llamada sobrevivió para indicar si el paciente estaba vivo o no al final del estudio.30

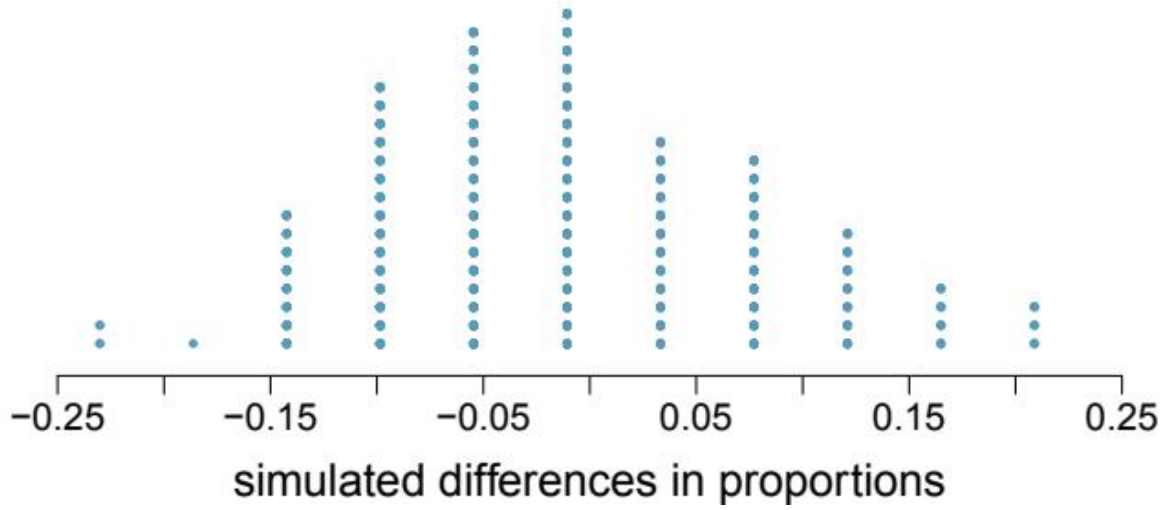


- (a) Según el diagrama de mosaico, ¿la supervivencia es independiente de si el paciente recibió o no un trasplante? Explique su razonamiento.
- (b) ¿Qué sugieren los diagramas de caja a continuación sobre la eficacia (efectividad) del tratamiento de trasplante de corazón?
- (c) ¿Qué proporción de pacientes en el grupo de tratamiento y qué proporción de pacientes en el grupo de control murieron?
- (d) Un enfoque para investigar si el tratamiento es efectivo o no es utilizar una técnica de aleatorización.
  - i. ¿Cuáles son las afirmaciones que se están probando?
  - ii. El párrafo a continuación describe la configuración de dicho enfoque, si tuviéramos que hacerlo sin utilizar software estadístico. Complete los espacios en blanco con un número o frase, lo que sea apropiado.

Escribimos vivo en tarjetas que representan a los pacientes que estaban vivos al final del estudio, y muerto en tarjetas que representan a los pacientes que no lo estaban. Luego, barajamos estas tarjetas y las dividimos en dos grupos: un grupo de tamaño que representa el tratamiento y otro grupo de tamaño que representa el control. Calculamos la diferencia entre la proporción de tarjetas de muertos en los grupos de tratamiento y control (tratamiento - control) y registramos este valor. Repetimos esto 100 veces para construir una distribución centrada en . Por último, calculamos la fracción de simulaciones donde las diferencias simuladas en proporciones son . Si esta fracción es baja, concluimos que es poco probable que hayamos

observado tal resultado por casualidad y que la hipótesis nula debe ser rechazada en favor de la alternativa.

iii. ¿Qué sugieren los resultados de la simulación que se muestran a continuación sobre la efectividad del programa de trasplante?



30B. Turnbull et al. "Survivorship of Heart Transplant Data". In: Journal of the American Statistical Association 69 (1974), pp. 74–80.

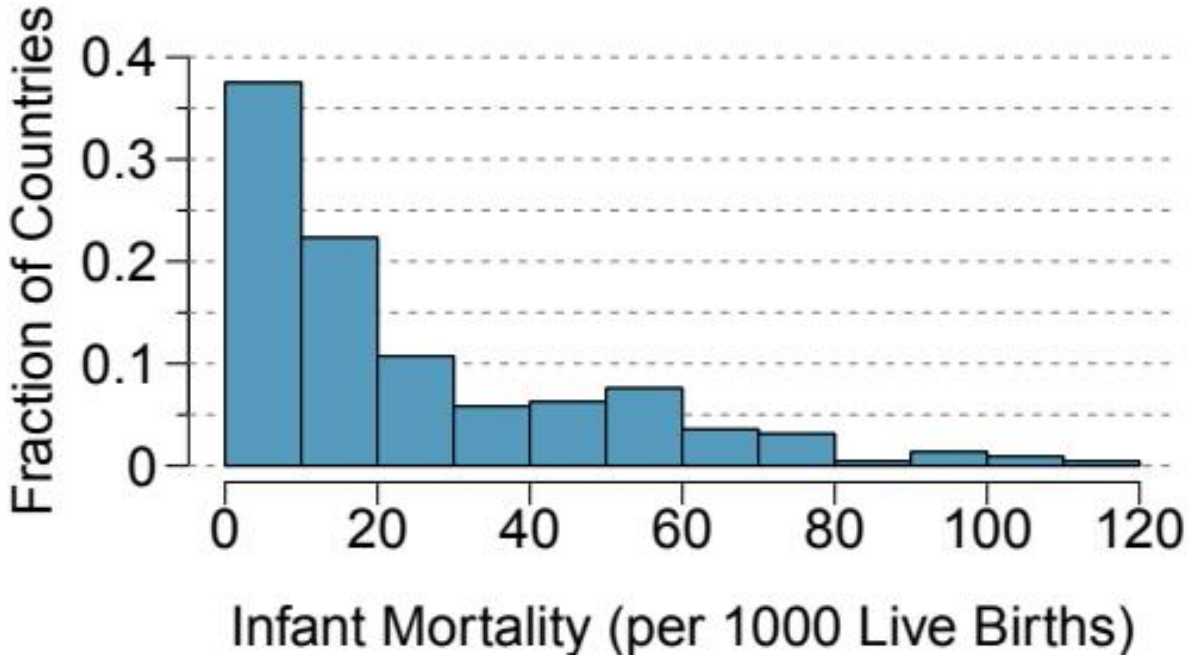
Ejercicios del Capítulo

2.27 Examen de recuperación. En una clase de 25 estudiantes, 24 de ellos hicieron un examen en clase y 1 estudiante hizo un examen de recuperación al día siguiente. El profesor calificó el primer lote de 24 exámenes y encontró una puntuación media de 74 puntos con una desviación estándar de 8.9 puntos. El estudiante que hizo la recuperación al día siguiente obtuvo 64 puntos en el examen.

- (a) ¿La puntuación del nuevo estudiante aumenta o disminuye la puntuación media?
- (b) ¿Cuál es la nueva media?
- (c) ¿La puntuación del nuevo estudiante aumenta o disminuye la desviación estándar de las puntuaciones?

2.28 Mortalidad infantil. La tasa de mortalidad infantil se define como el número de muertes infantiles por cada 1.000 nacidos vivos. Esta tasa se utiliza a menudo como un indicador del nivel de salud en un país. El histograma de frecuencia relativa que aparece a continuación muestra la distribución de las tasas de mortalidad infantil estimadas para 224 países para los que se disponía de dichos datos en 2014.31

- (a) Estimar Q1, la mediana y Q3 a partir del histograma.
- (b) ¿Esperaría que la media de este conjunto de datos fuera menor o mayor que la mediana? Explique su razonamiento.



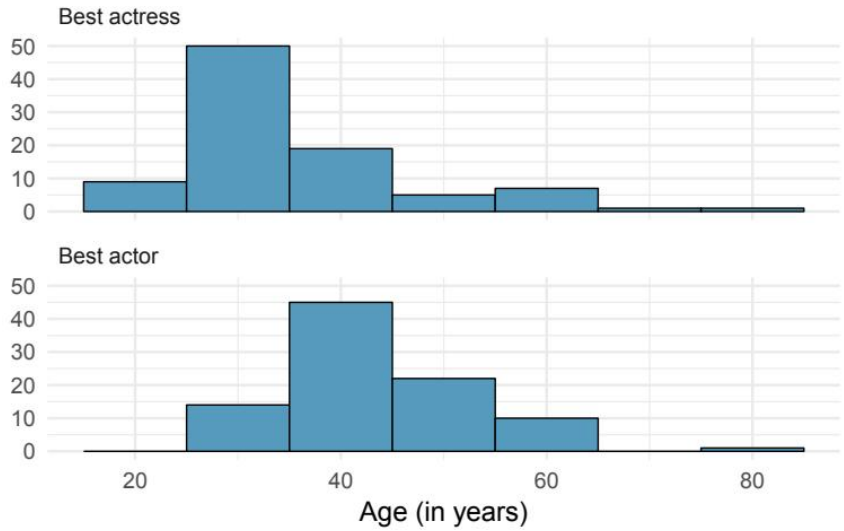
2.29 Espectadores de televisión. A los estudiantes de una clase de Estadística AP se les preguntó cuántas horas de televisión ven por semana (incluyendo la transmisión en línea). Esta muestra arrojó una media de 4.71 horas, con una desviación estándar de 4.18 horas. ¿Es simétrica la distribución del número de horas que los estudiantes ven la televisión semanalmente? Si no, ¿qué forma esperaría que tuviera esta distribución? Explique su razonamiento.

2.30 Una nueva estadística. La estadística $\bar{x}$ mediana puede utilizarse como una medida de asimetría. Supongamos que tenemos una distribución donde todas las observaciones son mayores que 0, xi > 0. ¿Cuál es la forma esperada de la distribución bajo las siguientes condiciones? Explique su razonamiento.

(a) $\frac{x}{mediana} = 1$
(b) $\frac{x}{mediana} < 1$
(c) $\frac{x}{mediana} > 1$

2.31 Ganadores del Oscar. Los primeros premios Oscar al mejor actor y a la mejor actriz se entregaron en 1929. Los histogramas siguientes muestran la distribución de edades de todos los ganadores del Oscar al mejor actor y a la mejor actriz desde 1929 hasta 2018. También se

proporcionan estadísticas resumidas para estas distribuciones. Compare las distribuciones de edades de los ganadores al mejor actor y a la mejor actriz.32



	Best Actress
Mean	36.2
SD	11.9
n	92

	Best Actor
Mean	43.8
SD	8.83
n	92

31CIA Factbook, Comparaciones entre países, 2014.

32Ganadores del Oscar de 1929 a 2012, datos hasta 2009 del archivo de datos del Journal of Statistics Education y datos más actuales de wikipedia.org.

2.32 Puntuaciones de los exámenes. La media en un examen de historia (puntuado sobre 100 puntos) fue de 85, con una desviación estándar de 15. ¿Es simétrica la distribución de las puntuaciones en este examen? Si no, ¿qué forma esperaría que tuviera esta distribución? Explique su razonamiento.

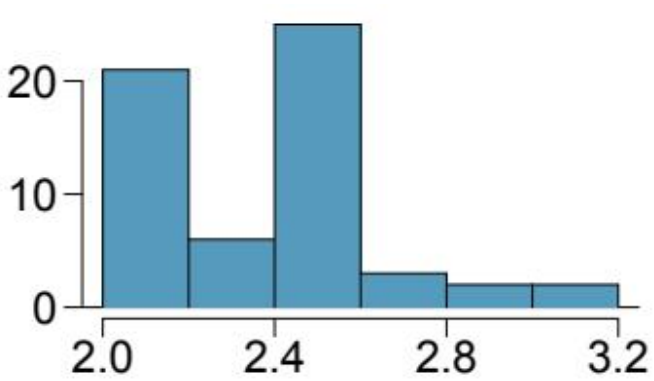
2.33 Puntuaciones de estadística. A continuación, se muestran las puntuaciones del examen final de veinte estudiantes de estadística introductoria.

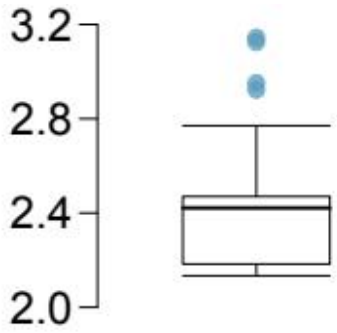
$$157, 66, 69, 71, 72, 73, 74, 77, 78, 79, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94$$

Crea un diagrama de caja de la distribución de estas puntuaciones. El resumen de cinco números que se proporciona a continuación puede ser útil.

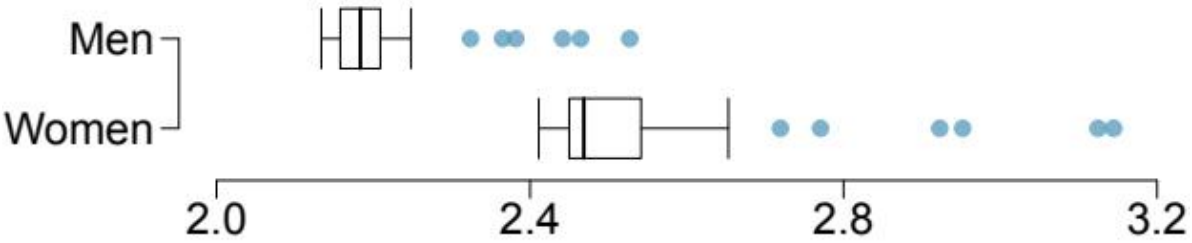
Min	Q1	Q2 (Mediana)	Q3	Max
57	72.5	78.5	82.5	94

2.34 Ganadores de maratones. El histograma y los diagramas de caja que se muestran a continuación muestran la distribución de los tiempos de finalización en horas para los ganadores masculinos y femeninos de la Maratón de Nueva York entre 1970 y 1999.

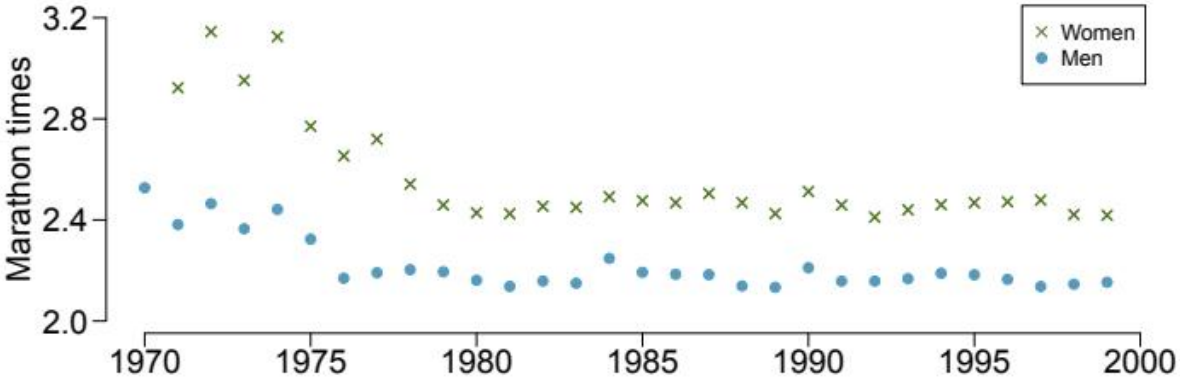




- (a) ¿Qué características de la distribución son evidentes en el histograma y no en el diagrama de caja? ¿Qué características son evidentes en el diagrama de caja pero no en el histograma?
- (b) ¿Cuál puede ser la razón de la distribución bimodal? Explique.
- (c) Compare la distribución de los tiempos de maratón para hombres y mujeres basándose en el diagrama de caja que se muestra a continuación.



(d) El diagrama de series temporales que se muestra a continuación es otra forma de ver estos datos. Describa lo que es visible en este diagrama pero no en los demás.



Capítulo 3

79

Probabilidad

- **3.1 Definiendo probabilidad**
- **3.2 Probabilidad condicional**
- **3.3 Muestreo de una población pequeña**
- **3.4 Variables aleatorias**
- **3.5 Distribuciones continuas**

La probabilidad forma la base de la estadística, y probablemente ya conozcas muchas de las ideas presentadas en este capítulo. Sin embargo, la formalización de los conceptos de probabilidad es probablemente nueva para la mayoría de los lectores.

Si bien este capítulo proporciona una base teórica para las ideas en capítulos posteriores y proporciona un camino hacia una comprensión más profunda, no se requiere el dominio de los conceptos introducidos en este capítulo para aplicar los métodos introducidos en el resto de este libro.

Para videos, diapositivas y otros recursos, por favor visita www.openintro.org/os

3.1 Definiendo probabilidad

La estadística se basa en la probabilidad, y aunque la probabilidad no es necesaria para las técnicas aplicadas en este libro, puede ayudarte a obtener una comprensión más profunda de los métodos y establecer una mejor base para cursos futuros.

3.1.1 Ejemplos introductorios

Antes de entrar en ideas técnicas, veamos algunos ejemplos básicos que pueden resultar más familiares.

EJEMPLO 3.1

Un "dado" (la forma singular de "dados") es un cubo con seis caras numeradas 1, 2, 3, 4, 5 y 6. ¿Cuál es la probabilidad de obtener un 1 al tirar un dado?

Si el dado es justo, entonces la probabilidad de obtener un 1 es tan buena como la probabilidad de cualquier otro número. Dado que hay seis resultados, la probabilidad debe ser 1 entre 6 o, equivalentemente, 1/6.

EJEMPLO 3.2

¿Cuál es la probabilidad de obtener un 1 o un 2 en la próxima tirada?

1 y 2 constituyen dos de los seis resultados posibles igualmente probables, por lo que la probabilidad de obtener uno de estos dos resultados debe ser $2/6 = 1/3$.

EJEMPLO 3.3

¿Cuál es la probabilidad de obtener 1, 2, 3, 4, 5 o 6 en la próxima tirada?

100%. El resultado debe ser uno de estos números.

EJEMPLO 3.4

¿Cuál es la probabilidad de no sacar un 2?

Dado que la probabilidad de sacar un 2 es 1/6 o $16.\bar{6}\%$, la probabilidad de no sacar un 2 debe ser $100\% - 16.\bar{6}\% = 83.\bar{3}\%$ o 5/6.

Alternativamente, podríamos haber notado que no sacar un 2 es lo mismo que obtener un 1, 3, 4, 5 o 6, lo que constituye cinco de los seis resultados igualmente probables y tiene una probabilidad de 5/6.

EJEMPLO 3.5

Consideremos el lanzamiento de dos dados. Si 1/6 de las veces el primer dado es un 1 y 1/6 de esas veces el segundo dado es un 1, ¿cuál es la probabilidad de obtener dos 1s?

Si el 16.$\bar{}$6% de las veces el primer dado es un 1 y 1/6 de esas veces el segundo dado también es un 1, entonces la probabilidad de que ambos dados sean 1 es (1/6) × (1/6) o 1/36.

81

3.1.2 Probabilidad

Utilizamos la probabilidad para construir herramientas que describan y comprendan la aleatoriedad aparente. A menudo enmarcamos la probabilidad en términos de un proceso aleatorio que da lugar a un resultado.

$$\begin{aligned} \text{Lanzar un dado} &\rightarrow \mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{4}, \mathbf{5}, \text{ o } \mathbf{6} \\ \text{Lanzar una moneda} &\rightarrow \mathbf{H} \text{ o } \mathbf{T} \end{aligned}$$

Lanzar un dado o una moneda es un proceso aparentemente aleatorio y cada uno da lugar a un resultado.

PROBABILIDAD

La probabilidad de un resultado es la proporción de veces que ocurriría el resultado si observáramos el proceso aleatorio un número infinito de veces.

La probabilidad se define como una proporción y siempre toma valores entre 0 y 1 (inclusive). También puede mostrarse como un porcentaje entre 0% y 100%.

La probabilidad puede ilustrarse lanzando un dado muchas veces. Sea ˆpn la proporción de resultados que son 1 después de los primeros n lanzamientos. A medida que aumenta el número de lanzamientos, ˆpn convergerá a la probabilidad de lanzar un 1, p = 1/6. La Figura 3.1 muestra esta convergencia para 100,000 lanzamientos de dados. La tendencia de ˆpn a estabilizarse alrededor de p se describe mediante la Ley de los Grandes Números.

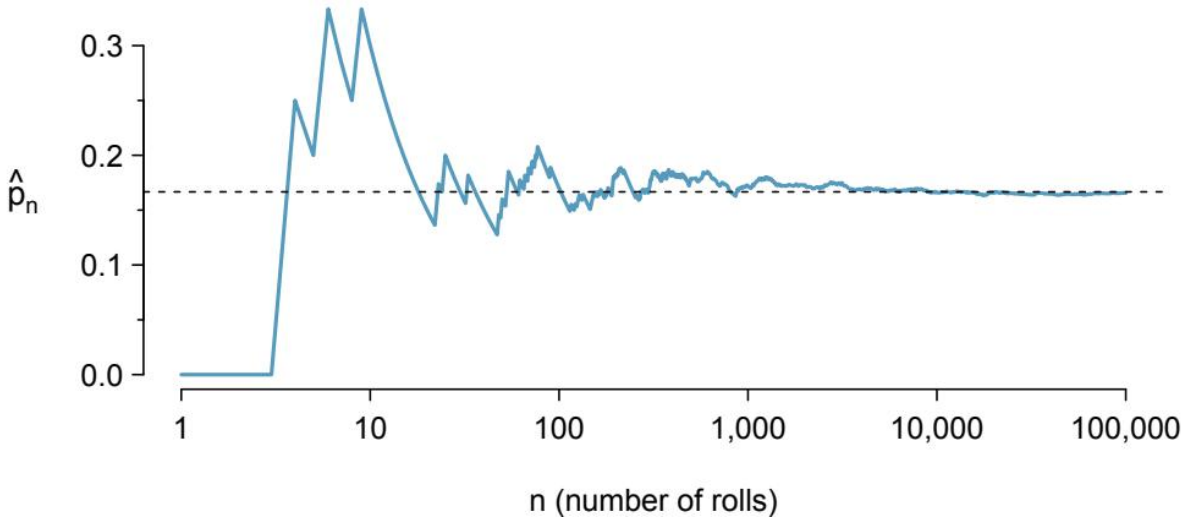


Figura 3.1: La fracción de lanzamientos de dados que son 1 en cada etapa de una simulación. La proporción tiende a acercarse a la probabilidad 1/6 0.167 a medida que aumenta el número de lanzamientos.

LEY DE LOS GRANDES NÚMEROS

A medida que se recopilan más observaciones, la proporción ˆpn de ocurrencias con un resultado particular converge a la probabilidad p de ese resultado.

Ocasionalmente, la proporción se desviará de la probabilidad y parecerá desafiar la Ley de los Grandes Números, como ˆpn lo hace muchas veces en la Figura 3.1. Sin embargo, estas desviaciones se hacen más pequeñas a medida que aumenta el número de lanzamientos.

Arriba escribimos p como la probabilidad de lanzar un 1. También podemos escribir esta probabilidad como

P(lanzar un 1)

A medida que nos sintamos más cómodos con esta notación, la abreviaremos aún más. Por ejemplo, si está claro que el proceso es “lanzar un dado”, podríamos abreviar P(lanzar un 1) como P(1).

PRÁCTICA GUIADA 3.6

Los procesos aleatorios incluyen lanzar un dado y lanzar una moneda. (a) Piensa en otro proceso aleatorio. (b) Describe todos los posibles resultados de ese proceso. Por ejemplo, lanzar un dado es un proceso aleatorio con posibles resultados 1, 2, …, 6. 1

Lo que pensamos como procesos aleatorios no son necesariamente aleatorios, sino que pueden ser simplemente demasiado difíciles de entender con exactitud. El cuarto ejemplo en la solución de la nota al pie de la Práctica Guiada 3.6 sugiere que el comportamiento de un compañero de cuarto es un proceso aleatorio. Sin embargo, incluso si el comportamiento de un compañero de cuarto no es verdaderamente aleatorio, modelar su comportamiento como un proceso aleatorio aún puede ser útil.

3.1.3 Resultados Disjuntos o Mutuamente Excluyentes

Dos resultados se denominan disjuntos o mutuamente excluyentes si no pueden ocurrir ambos. Por ejemplo, si lanzamos un dado, los resultados 1 y 2 son disjuntos ya que no pueden ocurrir ambos. Por otro lado, los resultados 1 y "sacar un número impar" no son disjuntos ya que ambos ocurren si el resultado del lanzamiento es un 1. Los términos disjuntos y mutuamente excluyentes son equivalentes e intercambiables.

Calcular la probabilidad de resultados disjuntos es fácil. Al lanzar un dado, los resultados 1 y 2 son disjuntos, y calculamos la probabilidad de que uno de estos resultados ocurra sumando sus probabilidades separadas:

$$P(\mathbf{1} \text{ o } \mathbf{2}) = P(\mathbf{1}) + P(\mathbf{2}) = 1/6 + 1/6 = 1/3$$

¿Qué pasa con la probabilidad de sacar un 1, 2, 3, 4, 5 o 6? Aquí nuevamente, todos los resultados son disjuntos, por lo que sumamos las probabilidades:

$$\begin{aligned} &P(\mathbf{1} \text{ o } \mathbf{2} \text{ o } \mathbf{3} \text{ o } \mathbf{4} \text{ o } \mathbf{5} \text{ o } \mathbf{6}) \\ &= P(\mathbf{1}) + P(\mathbf{2}) + P(\mathbf{3}) + P(\mathbf{4}) + P(\mathbf{5}) + P(\mathbf{6}) \\ &= 1/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 1 \end{aligned}$$

La Regla de la Adición garantiza la precisión de este enfoque cuando los resultados son disjuntos.

REGLA DE LA ADICIÓN DE RESULTADOS DISJUNTOS

Si A1 y A2 representan dos resultados disjuntos, entonces la probabilidad de que uno de ellos ocurra está dada por

$$P(A\_1 \text{ o } A\_2) = P(A\_1) + P(A\_2)$$

Si hay muchos resultados disjuntos A1, …, Ak, entonces la probabilidad de que uno de estos resultados ocurra es

$$P(A\_1) + P(A\_2) + \cdots + P(A\_k)$$

1Aquí hay cuatro ejemplos. (i) Si alguien se enferma o no en el próximo mes es un proceso aparentemente aleatorio con los resultados enfermo y no. (ii) Podemos generar un proceso aleatorio seleccionando aleatoriamente a una persona y midiendo la altura de esa persona. El resultado de este proceso será un número positivo. (iii) Si el mercado de valores sube o baja la semana que viene es un proceso aparentemente aleatorio con los posibles resultados de subida, bajada y sin cambios. Alternativamente, podríamos haber utilizado el cambio porcentual en el mercado de valores como un resultado numérico. (iv) Si tu compañero de cuarto lava sus platos esta noche probablemente parece un proceso aleatorio con los posibles resultados de lava los platos y deja los platos.

Estamos interesados en la probabilidad de sacar un 1, 4 o 5. (a) Explica por qué los resultados 1, 4 y 5 son disjuntos. (b) Aplica la Regla de la Adición para resultados disjuntos para determinar P(1 o 4 o 5).2

PRÁCTICA GUIADA 3.8

En el conjunto de datos de préstamos del Capítulo 2, la variable de propiedad de la vivienda describía si el prestatario alquila, tiene una hipoteca o es dueño de su propiedad. De los 10,000 prestatarios, 3858 alquilaron, 4789 tenían una hipoteca y 1353 eran dueños de su casa.3

- (a) ¿Son los resultados alquiler, hipoteca y propiedad disjuntos?
- (b) Determina la proporción de préstamos con valor hipoteca y propiedad por separado.
- (c) Utiliza la Regla de la Adición para resultados disjuntos para calcular la probabilidad de que un préstamo seleccionado aleatoriamente del conjunto de datos sea para alguien que tiene una hipoteca o es dueño de su casa.

Los científicos de datos rara vez trabajan con resultados individuales y en su lugar consideran conjuntos o colecciones de resultados. Sea A el evento donde un lanzamiento de dado resulta en 1 o 2 y B representa el evento de que el lanzamiento de dado sea un 4 o un 6. Escribimos A como el conjunto de resultados $\{1, 2\}$ y B = $\{4, 6\}$. Estos conjuntos se denominan comúnmente eventos. Debido a que A y B no tienen elementos en común, son eventos disjuntos. A y B se representan en la Figura 3.2.

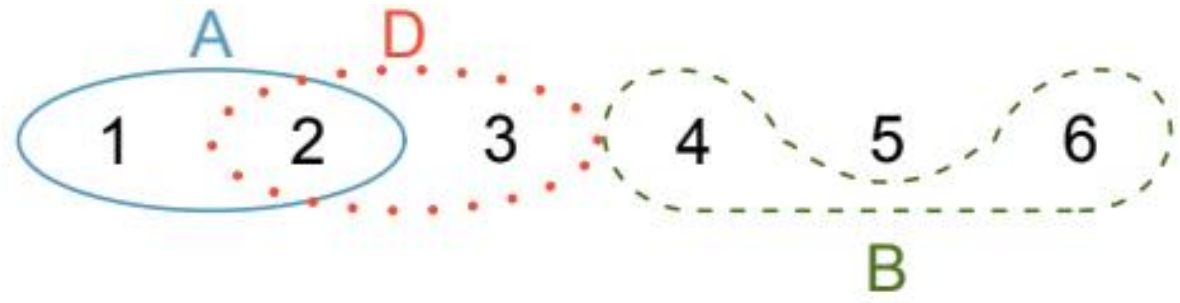


Figura 3.2: Tres eventos, A, B y D, consisten en resultados de lanzar un dado. A y B son disjuntos ya que no tienen ningún resultado en común.

La Regla de la Adición se aplica tanto a resultados disjuntos como a eventos disjuntos. La probabilidad de que uno de los eventos disjuntos A o B ocurra es la suma de las probabilidades separadas:

$$P(A \text{ o } B) = P(A) + P(B) = 1/3 + 1/3 = 2/3$$

PRÁCTICA GUIADA 3.9

(a) Verifica que la probabilidad del evento A, P(A), es 1/3 utilizando la Regla de la Adición. (b) Haz lo mismo para el evento B. 4

PRÁCTICA GUIADA 3.10

(a) Usando la Figura 3.2 como referencia, ¿qué resultados están representados por el evento D? (b) ¿Son los eventos B y D disjuntos? (c) ¿Son los eventos A y D disjuntos?5

PRÁCTICA GUIADA 3.11

En la Práctica Guiada 3.10, confirmaste que B y D de la Figura 3.2 son disjuntos. Calcula la probabilidad de que ocurra el evento B o el evento D.6

6Dado que B y D son eventos disjuntos, usa la Regla de la Adición: P(B o D) = P(B) + P(D) = 1 3 + 1 3 = 2 3 .

2 (a) El proceso aleatorio es lanzar un dado, y a lo sumo uno de estos resultados puede salir. Esto significa que son resultados disjuntos. (b) P(1 o 4 o 5) = P(1) + P(4) + P(5) = 1 6 + 1 6 + 1 6 = 3 6 = 1 2

3 (a) Sí. Cada préstamo se clasifica en solo un nivel de propiedad de la vivienda. (b) Hipoteca: 4789 10000 = 0.479. Propia: 1353 10000 = 0.135. (c) P(hipoteca o propia) = P(hipoteca) + P(propia) = 0.479 + 0.135 = 0.614.

4 (a) P(A) = P(1 o 2) = P(1) + P(2) = 1 6 + 1 6 = 2 6 = 1 3 . (b) De manera similar, P(B) = 1/3.

5 (a) Resultados 2 y 3. (b) Sí, los eventos B y D son disjuntos porque no comparten resultados. (c) Los eventos A y D comparten un resultado en común, 2, por lo que no son disjuntos.

3.1.4 Probabilidades cuando los eventos no son disjuntos

Consideremos los cálculos para dos eventos que no son disjuntos en el contexto de una baraja regular de 52 cartas, representada en la Figura 3.3. Si no estás familiarizado con las cartas de una baraja regular, consulta la nota al pie.7

2	3	4	5	6	7	8	9	10	J	Q	K	A
2	3	4	5	6	7	8	9	10	J	Q	K	A
2	3	4	5	6	7	8	9	10	J	Q	K	A
2	3	4	5	6	7	8	9	10	J	Q	K	A

Figura 3.3: Representaciones de las 52 cartas únicas en una baraja.

PRÁCTICA GUIADA 3.12

(a) ¿Cuál es la probabilidad de que una carta seleccionada al azar sea un diamante? (b) ¿Cuál es la probabilidad de que una carta seleccionada al azar sea una figura?8

Los diagramas de Venn son útiles cuando los resultados se pueden clasificar como “dentro” o “fuera” para dos o tres variables, atributos o procesos aleatorios. El diagrama de Venn en la Figura 3.4 usa un círculo para representar los diamantes y otro para representar las figuras. Si una carta es tanto un diamante como una figura, cae en la intersección de los círculos. Si es un diamante pero no una figura, estará en parte del círculo izquierdo que no está en el círculo derecho (y así sucesivamente). El número total de cartas que son diamantes viene dado por el número total de cartas en el círculo de diamantes: 10 + 3 = 13. También se muestran las probabilidades (por ejemplo, 10/52 = 0.1923).

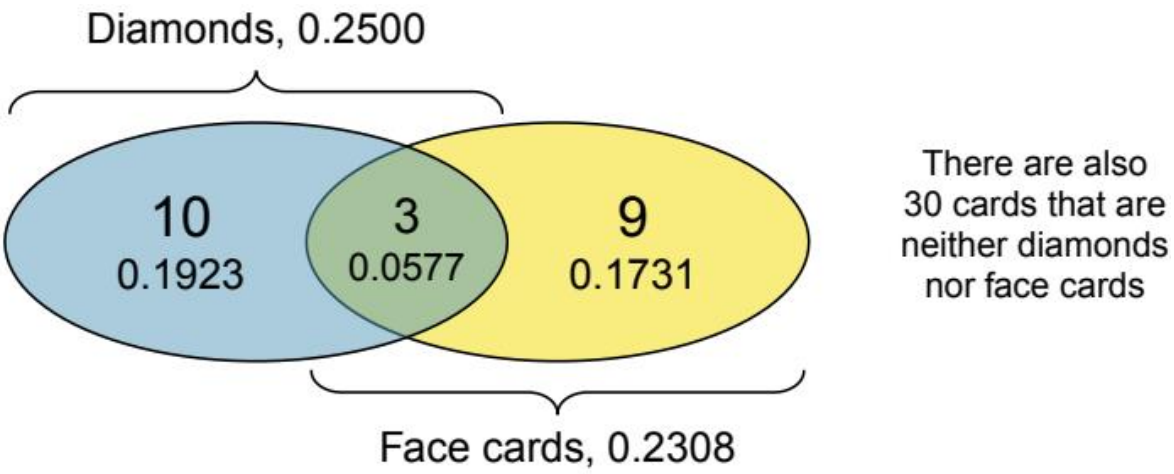


Figura 3.4: Un diagrama de Venn para diamantes y figuras.

Sea A el evento de que una carta seleccionada al azar sea un diamante y B el evento de que sea una figura. ¿Cómo calculamos P(A o B)? Los eventos A y B no son disjuntos (las cartas J , Q y K entran en ambas categorías), por lo que no podemos usar la Regla de la Adición para eventos disjuntos. En cambio, usamos el diagrama de Venn. Comenzamos sumando las probabilidades de los dos eventos:

$$P(A) + P(B) = P(\diamondsuit) + P(\text{figura}) = 13/52 + 12/52$$

7Las 52 cartas se dividen en cuatro palos: (trébol), (diamante), (corazón), (pica). Cada palo tiene sus 13 cartas etiquetadas: 2, 3, ..., 10, J (jota), Q (reina), K (rey) y A (as). Por lo tanto, cada carta es una combinación única de un palo y una etiqueta, por ejemplo, 4 y J . Las 12 cartas representadas por las jotas, reinas y reyes se llaman figuras. Las cartas que son o suelen ser de color rojo, mientras que los otros dos palos suelen ser de color negro.

8 (a) Hay 52 cartas y 13 diamantes. Si las cartas están bien barajadas, cada carta tiene la misma probabilidad de ser extraída, por lo que la probabilidad de que una carta seleccionada al azar sea un diamante es P( ) = 13 52 = 0.250. (b) Del mismo modo, hay 12 figuras, por lo que P(figura) = 12 52 = 3 13 = 0.231.

Sin embargo, las tres cartas que están en ambos eventos se contaron dos veces, una vez en cada probabilidad. Debemos corregir este conteo doble:

$$\begin{aligned} P(A \text{ o } B) &= P(\diamond \text{ o figura}) \\ &= P(\diamond) + P(\text{figura}) - P(\diamond \text{ y figura}) \\ &= 13/52 + 12/52 - 3/52 \\ &= 22/52 = 11/26 \end{aligned}$$

Esta ecuación es un ejemplo de la Regla General de la Adición.

REGLA GENERAL DE LA ADICIÓN

Si A y B son dos eventos cualesquiera, disjuntos o no, entonces la probabilidad de que al menos uno de ellos ocurra es

$$P(A \text{ o } B) = P(A) + P(B) - P(A \text{ y } B)$$

donde P(A y B) es la probabilidad de que ocurran ambos eventos.

CONSEJO: "o" es inclusivo

Cuando escribimos "o" en estadística, nos referimos a "y/o" a menos que explícitamente digamos lo contrario. Por lo tanto, A o B ocurre significa que A, B, o ambos A y B ocurren.

PRÁCTICA GUIADA 3.13

(a) Si A y B son disjuntos, describa por qué esto implica que P(A y B) = 0. (b) Usando la parte (a), verifique que la Regla General de la Adición se simplifica a la Regla de Adición más simple para eventos disjuntos si A y B son disjuntos.9

PRÁCTICA GUIADA 3.14

En el conjunto de datos de préstamos que describe 10,000 préstamos, 1495 préstamos fueron de solicitudes conjuntas (por ejemplo, una pareja solicitó junta), 4789 solicitantes tenían una hipoteca y 950 tenían ambas características. Cree un diagrama de Venn para esta configuración.10

PRÁCTICA GUIADA 3.15

(a) Utilice su diagrama de Venn de la Práctica Guiada 3.14 para determinar la probabilidad de que un préstamo extraído aleatoriamente del conjunto de datos de préstamos sea de una solicitud conjunta donde la pareja tenía una hipoteca. (b) ¿Cuál es la probabilidad de que el préstamo tenga alguno de estos atributos?11

10Tanto los conteos como las probabilidades correspondientes (p. ej., 3839/10000 = 0.384) se muestran. Observe que el número de préstamos representados en el círculo izquierdo corresponde a 3839 + 950 = 4789, y el número representado en el círculo derecho es 950 + 545 = 1495.

11(a) La solución está representada por la intersección de los dos círculos: 0.095. (b) Esta es la suma de las tres probabilidades disjuntas que se muestran en los círculos: 0.384 + 0.095 + 0.055 = 0.534 (desviación de 0.001 debido a un error de redondeo).

9 (a) Si A y B son disjuntos, A y B nunca pueden ocurrir simultáneamente. (b) Si A y B son disjuntos, entonces el último término P(A y B) en la fórmula de la Regla General de Adición es 0 (ver parte (a)) y nos quedamos con la Regla de Adición para eventos disjuntos.

3.1.5 Distribuciones de probabilidad

Una distribución de probabilidad es una tabla de todos los resultados disjuntos y sus probabilidades asociadas. La Figura 3.5 muestra la distribución de probabilidad para la suma de dos dados.

Suma de los dados	2	3	4	5	6	7	8	9	10	11	12
Probabilidad	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Figura 3.5: Distribución de probabilidad para la suma de dos dados.

REGLAS PARA DISTRIBUCIONES DE PROBABILIDAD

Una distribución de probabilidad es una lista de los posibles resultados con las probabilidades correspondientes que satisfacen tres reglas:

- 1. Los resultados enumerados deben ser disjuntos.
- 2. Cada probabilidad debe estar entre 0 y 1.
- 3. Las probabilidades deben sumar 1.

PRÁCTICA GUIADA 3.16

La Figura 3.6 sugiere tres distribuciones para los ingresos familiares en los Estados Unidos. Sólo una es correcta. ¿Cuál debe ser? ¿Qué tienen de malo las otras dos?12

Rango de ingresos	$0-25k	$25k-50k	$50k-100k	$100k+
(a)	0.18	0.39	0.33	0.16
(b)	0.38	-0.27	0.52	0.37
(c)	0.28	0.27	0.29	0.16

Figura 3.6: Distribuciones propuestas de los ingresos familiares en EE. UU. (Práctica Guiada 3.16).

El Capítulo 1 enfatizó la importancia de graficar los datos para proporcionar resúmenes rápidos. Las distribuciones de probabilidad también se pueden resumir en un diagrama de barras. Por ejemplo, la distribución de los ingresos familiares en EE. UU. se muestra en la Figura 3.7 como un diagrama de barras. La distribución de probabilidad para la suma de dos dados se muestra en la Figura 3.5 y se grafica en la Figura 3.8.

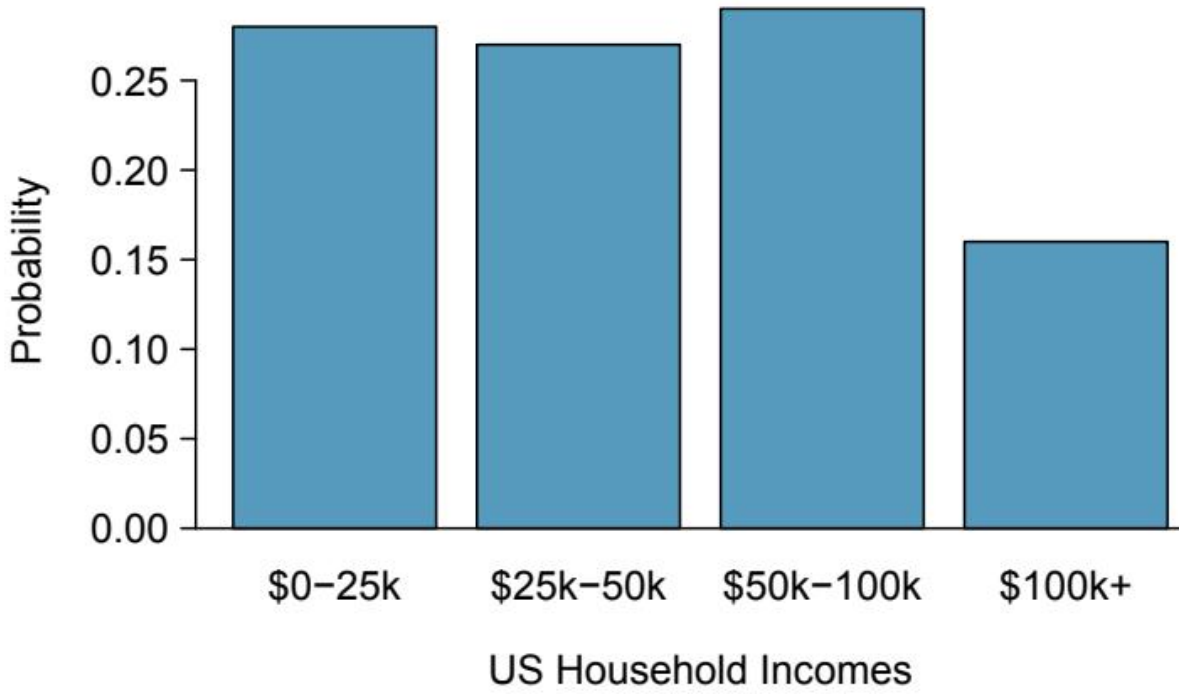


Figura 3.7: La distribución de probabilidad de los ingresos familiares en EE. UU.

12Las probabilidades de (a) no suman 1. La segunda probabilidad en (b) es negativa. Esto deja (c), que seguramente satisface los requisitos de una distribución. Se dijo que una de las tres era la distribución real de los ingresos familiares en EE. UU., por lo que debe ser (c).

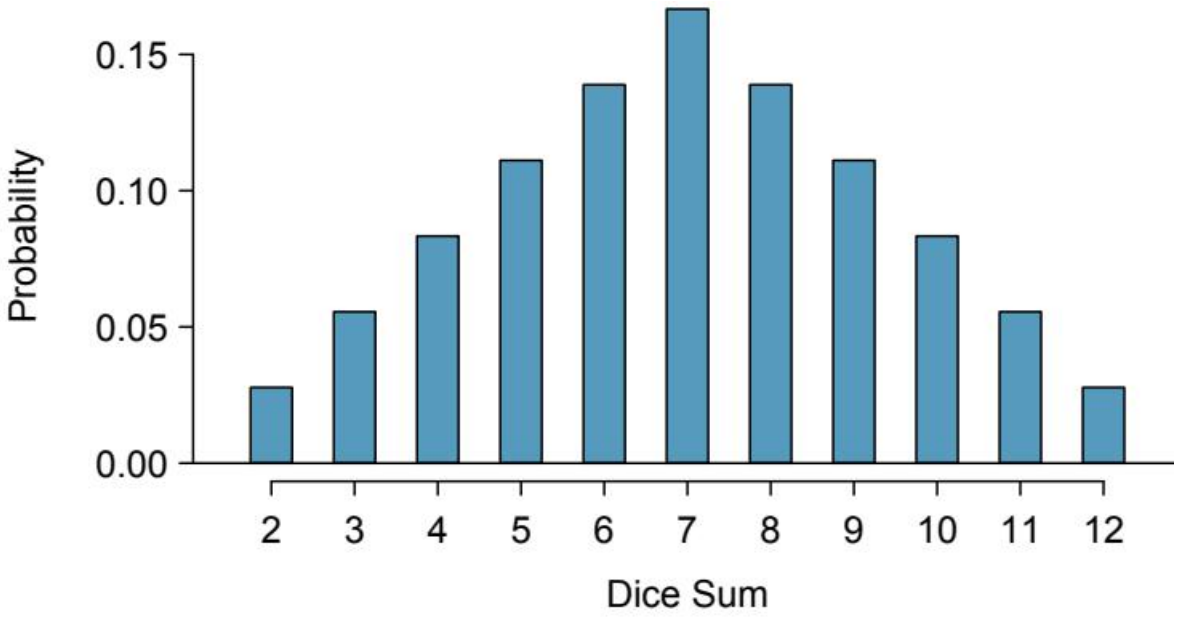


Figura 3.8: La distribución de probabilidad de la suma de dos dados.

En estos diagramas de barras, las alturas de las barras representan las probabilidades de los resultados. Si los resultados son numéricos y discretos, suele ser (visualmente) conveniente hacer un diagrama de barras que se asemeje a un histograma, como en el caso de la suma de dos dados. Otro ejemplo de trazar las barras en sus respectivas ubicaciones se muestra en la Figura 3.18 en la página 115.

3.1.6 Complemento de un evento

Lanzar un dado produce un valor en el conjunto {1, 2, 3, 4, 5, 6}. Este conjunto de todos los resultados posibles se llama espacio muestral (S) para lanzar un dado. A menudo utilizamos el espacio muestral para examinar el escenario donde un evento no ocurre.

Sea D = {2, 3} el evento en que el resultado de lanzar un dado es 2 o 3. Entonces, el complemento de D representa todos los resultados en nuestro espacio muestral que no están en D, lo cual se denota como Dc = {1, 4, 5, 6}. Es decir, Dc es el conjunto de todos los resultados posibles no incluidos ya en D. La Figura 3.9 muestra la relación entre D, Dc y el espacio muestral S.

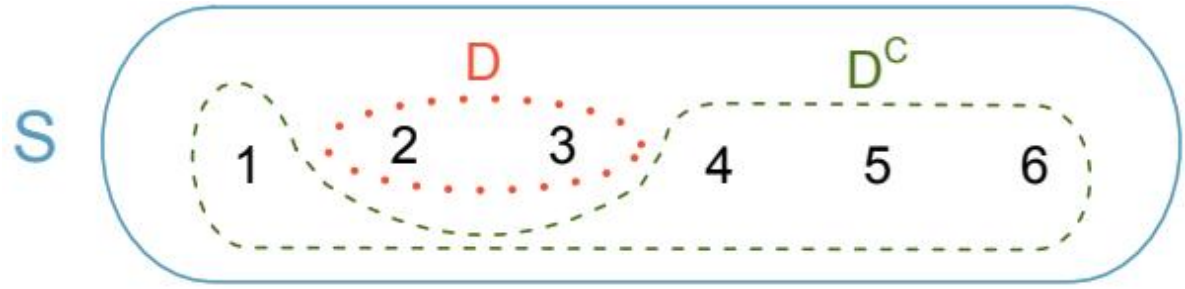


Figura 3.9: Evento D = {2, 3} y su complemento, Dc = {1, 4, 5, 6}. S representa el espacio muestral, que es el conjunto de todos los resultados posibles.

PRÁCTICA GUIADA 3.17

(a) Calcule P(Dc) = P(sacar un 1, 4, 5 o 6). (b) ¿Cuál es P(D) + P(Dc)?13

PRÁCTICA GUIADA 3.18

Los eventos A = {1, 2} y B = {4, 6} se muestran en la Figura 3.2 en la página 84. (a) Escriba lo que representan Ac y Bc . (b) Calcule P(Ac) y P(Bc). (c) Calcule P(A) + P(Ac) y P(B) + P(Bc).14

13(a) Los resultados son disjuntos y cada uno tiene una probabilidad de 1/6, por lo que la probabilidad total es 4/6 = 2/3. (b) También podemos ver que P(D) = 1 6 + 1 6 = 1/3. Dado que D y Dc son disjuntos, P(D) + P(Dc) = 1.

14Soluciones breves: (a) Ac = {3, 4, 5, 6} y Bc = {1, 2, 3, 5}. (b) Teniendo en cuenta que cada resultado es disjunto, sume las probabilidades de resultados individuales para obtener P(Ac) = 2/3 y P(Bc) = 2/3. (c) A y Ac son disjuntos, y lo mismo ocurre con B y Bc . Por lo tanto, P(A) + P(Ac) = 1 y P(B) + P(Bc) = 1.

3.1. DEFINICIÓN DE PROBABILIDAD 89

El complemento de un evento A se construye para tener dos propiedades muy importantes: (i) todo resultado posible que no esté en A está en Ac , y (ii) A y Ac son disjuntos. La propiedad (i) implica

$$P(A \text{ o } A^c) = 1$$

Es decir, si el resultado no está en A, debe estar representado en Ac . Usamos la Regla de Adición para eventos disjuntos para aplicar la Propiedad (ii):

$$P(A \text{ o } A^c) = P(A) + P(A^c)$$

La combinación de las dos últimas ecuaciones produce una relación muy útil entre la probabilidad de un evento y su complemento.

COMPLEMENTO

El complemento del evento A se denota Ac , y Ac representa todos los resultados que no están en A. A y Ac están relacionados matemáticamente:

$$P(A) + P(A^c) = 1, \quad \text{es decir.} \quad P(A) = 1 - P(A^c)$$

En ejemplos simples, calcular A o Ac es factible en unos pocos pasos. Sin embargo, usar el complemento puede ahorrar mucho tiempo a medida que los problemas crecen en complejidad.

PRÁCTICA GUIADA 3.19

Sea A el evento en el que lanzamos dos dados y su total es menor que 12. (a) ¿Qué representa el evento Ac ? (b) Determine P(Ac) a partir de la Figura 3.5 en la página 87. (c) Determine P(A).15

PRÁCTICA GUIADA 3.20

Encuentre las siguientes probabilidades para lanzar dos dados:16

- (a) La suma de los dados no es 6.
- (b) La suma es al menos 4. Es decir, determine la probabilidad del evento B = {4, 5, …, 12}.
- (c) La suma no es mayor que 10. Es decir, determine la probabilidad del evento D = {2, 3, …, 10}.

3.1.7 Independencia

Así como las variables y las observaciones pueden ser independientes, los procesos aleatorios también pueden ser independientes. Dos procesos son independientes si conocer el resultado de uno no proporciona información útil sobre el resultado del otro. Por ejemplo, lanzar una moneda y tirar un dado son dos procesos independientes: saber que la moneda salió cara no ayuda a determinar el resultado de una tirada de dados. Por otro lado, los precios de las acciones generalmente suben o bajan juntos, por lo que no son independientes.

El Ejemplo 3.5 proporciona un ejemplo básico de dos procesos independientes: lanzar dos dados. Queremos determinar la probabilidad de que ambos sean 1. Suponga que uno de los dados es rojo y el otro blanco. Si el resultado del dado rojo es un 1, no proporciona información sobre el resultado del dado blanco. Primero encontramos esta misma pregunta en el Ejemplo 3.5 (página 81), donde calculamos la probabilidad utilizando el siguiente razonamiento: 1/6 del tiempo el dado rojo es un 1, y 1/6 de esas veces el dado blanco

15(a) El complemento de A: cuando el total es igual a 12. (b) P(Ac) = 1/36. (c) Use la probabilidad del complemento de la parte (b), P(Ac) = 1/36, y la ecuación para el complemento: P(menos de 12) = 1 − P(12) = 1 − 1/36 = 35/36.

16(a) Primero encuentre P(6) = 5/36, luego use el complemento: P(no 6) = 1 − P(6) = 31/36.

(b) Primero encuentre el complemento, que requiere mucho menos esfuerzo: P(2 o 3) = 1/36 + 2/36 = 1/12. Luego calcule P(B) = 1 − P(Bc) = 1 − 1/12 = 11/12.

(c) Como antes, encontrar el complemento es la forma inteligente de determinar P(D). Primero encuentre P(Dc) = P(11 o 12) = 2/36 + 1/36 = 1/12. Luego calcule P(D) = 1 − P(Dc) = 11/12.

también será 1. Esto se ilustra en la Figura 3.10. Debido a que los lanzamientos son independientes, las probabilidades de los resultados correspondientes se pueden multiplicar para obtener la respuesta final: (1/6)×(1/6) = 1/36. Esto se puede generalizar a muchos procesos independientes.

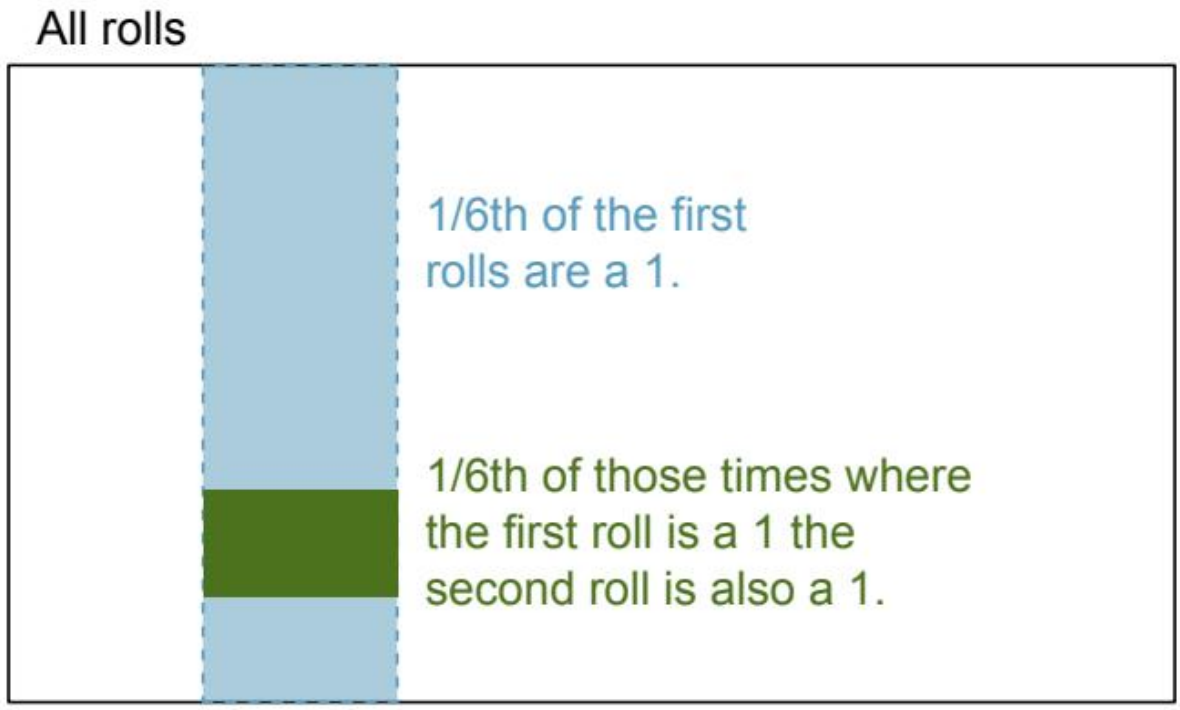


Figura 3.10: 1/6 del tiempo, el primer lanzamiento es un 1. Luego, 1/6 de esas veces, el segundo lanzamiento también será un 1.

EJEMPLO 3.21

¿Qué pasaría si también hubiera un dado azul independiente de los otros dos? ¿Cuál es la probabilidad de lanzar los tres dados y obtener todos 1s?

La misma lógica se aplica desde el Ejemplo 3.5. Si 1/36 del tiempo los dados blanco y rojo son ambos 1, entonces 1/6 de esas veces el dado azul también será 1, así que multiplique:

P(blanco = 1 y rojo = 1 y azul = 1) = P(blanco = 1) × P(rojo = 1) × P(azul = 1) = (1/6) × (1/6) × (1/6) = 1/216

El Ejemplo 3.21 ilustra lo que se llama la Regla de Multiplicación para procesos independientes.

REGLA DE MULTIPLICACIÓN PARA PROCESOS INDEPENDIENTES

Si A y B representan eventos de dos procesos diferentes e independientes, entonces la probabilidad de que ocurran tanto A como B se puede calcular como el producto de sus probabilidades separadas:

$$P(A \text{ y } B) = P(A) \times P(B)$$

De manera similar, si hay k eventos A1, …, Ak de k procesos independientes, entonces la probabilidad de que todos ocurran es

$$P(A\_1) \times P(A\_2) \times \cdots \times P(A\_k)$$

PRÁCTICA GUIADA 3.22

Aproximadamente el 9% de las personas son zurdas. Supongamos que se seleccionan 2 personas al azar de la población de EE. UU. Debido a que el tamaño de la muestra de 2 es muy pequeño en relación con la población, es razonable suponer que estas dos personas son independientes. (a) ¿Cuál es la probabilidad de que ambas sean zurdas? (b) ¿Cuál es la probabilidad de que ambas sean diestras?17

17(a) La probabilidad de que la primera persona sea zurda es 0.09, que es la misma para la segunda persona. Aplicamos la Regla de la Multiplicación para procesos independientes para determinar la probabilidad de que ambas sean zurdas: $0.09 \times 0.09 = 0.0081$.

(b) Es razonable suponer que la proporción de personas ambidiestras (tanto diestras como zurdas) es casi 0, lo que resulta en $\text{P(diestro)} = 1 - 0.09 = 0.91$. Usando el mismo razonamiento que en la parte (a), la probabilidad de que ambas sean diestras es $0.91 \times 0.91 = 0.8281$.

Supongamos que se seleccionan 5 personas al azar.18

- (a) ¿Cuál es la probabilidad de que todas sean diestras?
- (b) ¿Cuál es la probabilidad de que todas sean zurdas?
- (c) ¿Cuál es la probabilidad de que no todas las personas sean diestras?

Supongamos que las variables de lateralidad y sexo son independientes, es decir, saber el sexo de alguien no proporciona información útil sobre su lateralidad y viceversa. Entonces podemos calcular si una persona seleccionada al azar es diestra y mujer19 usando la Regla de la Multiplicación:

$$\begin{aligned} P(\text{diestro y mujer}) &= P(\text{diestro}) \times P(\text{mujer}) \\ &= 0.91 \times 0.50 = 0.455 \end{aligned}$$

PRÁCTICA GUIADA 3.24

Se seleccionan tres personas al azar.[20]

- (a) ¿Cuál es la probabilidad de que la primera persona sea hombre y diestro?
- (b) ¿Cuál es la probabilidad de que las dos primeras personas sean hombres y diestros?
- (c) ¿Cuál es la probabilidad de que la tercera persona sea mujer y zurda?
- (d) ¿Cuál es la probabilidad de que las dos primeras personas sean hombres y diestros y la tercera persona sea mujer y zurda?

A veces nos preguntamos si un resultado proporciona información útil sobre otro resultado. La pregunta que estamos haciendo es, ¿son independientes las ocurrencias de los dos eventos? Decimos que dos eventos A y B son independientes si satisfacen P(A y B) = P(A) × P(B).

EJEMPLO 3.25

Si mezclamos una baraja de cartas y sacamos una, ¿el evento de que la carta sea un corazón es independiente del evento de que la carta sea un as?

La probabilidad de que la carta sea un corazón es 1/4 y la probabilidad de que sea un as es 1/13. La probabilidad de que la carta sea el as de corazones es 1/52. Comprobamos si se cumple P(A y B) = P(A) × P(B):

$$P(\heartsuit) \times P(\text{as}) = \frac{1}{4} \times \frac{1}{13} = \frac{1}{52} = P(\heartsuit \text{ y as})$$

Debido a que la ecuación se cumple, el evento de que la carta sea un corazón y el evento de que la carta sea un as son eventos independientes.

P(los cinco son DD) = P(primero = DD, segundo = DD, …, quinto = DD)

= P(primero = DD) × P(segundo = DD) × · · · × P(quinto = DD)

- = 0.91 × 0.91 × 0.91 × 0.91 × 0.91 = 0.624
- (b) Usando el mismo razonamiento que en (a), 0.09 × 0.09 × 0.09 × 0.09 × 0.09 = 0.0000059
- (c) Usa el complemento, P(los cinco son DD), para responder esta pregunta:

P(no todos DD) = 1 − P(todos DD) = 1 − 0.624 = 0.376

[19]La proporción real de la población de EE. UU. que es femenina es aproximadamente del 50%, por lo que usamos 0.5 para la probabilidad de muestrear a una mujer. Sin embargo, esta probabilidad sí difiere en otros países.

[20]Se proporcionan respuestas breves. (a) Esto se puede escribir en notación de probabilidad como P(una persona seleccionada al azar es hombre y diestro) = 0.455. (b) 0.207. (c) 0.045. (d) 0.0093.

18(a) Las abreviaturas DD y ZD se utilizan para diestro y zurdo, respectivamente. Dado que cada uno es independiente, aplicamos la Regla de la Multiplicación para procesos independientes:

Ejercicios

- **3.1 Verdadero o falso.** Determine si las siguientes afirmaciones son verdaderas o falsas, y explique su razonamiento.
- (a) Si una moneda justa se lanza muchas veces y los últimos ocho lanzamientos son todos cara, entonces la probabilidad de que el siguiente lanzamiento sea cara es algo menor al 50%.
- (b) Sacar una figura (jota, reina o rey) y sacar una carta roja de una baraja completa de cartas son eventos mutuamente excluyentes.
- (c) Sacar una figura y sacar un as de una baraja completa de cartas son eventos mutuamente excluyentes.

3.2 Ruleta. El juego de la ruleta implica girar una rueda con 38 casillas: 18 rojas, 18 negras y 2 verdes. Una bola gira sobre la rueda y eventualmente aterriza en una casilla, donde cada casilla tiene la misma probabilidad de capturar la bola.

- (a) Usted observa una ruleta girar 3 veces consecutivas y la bola cae en una casilla roja cada vez. ¿Cuál es la probabilidad de que la bola caiga en una casilla roja en el siguiente giro?
- (b) Usted observa una ruleta girar 300 veces consecutivas y la bola cae en una casilla roja cada vez. ¿Cuál es la probabilidad de que la bola caiga en una casilla roja en el siguiente giro?
- (c) ¿Tiene la misma confianza en sus respuestas a las partes (a) y (b)? ¿Por qué o por qué no?

Foto de H ̊akan Dahlstr ̈om (http://flic.kr/p/93fEzp) Licencia CC BY 2.0

3.3 Cuatro juegos, un ganador. A continuación, se presentan cuatro versiones del mismo juego. Su archienemigo puede elegir la versión del juego, y luego usted puede elegir cuántas veces lanzar una moneda: 10 veces o 100 veces. Identifique cuántos lanzamientos de moneda debe elegir para cada versión del juego. Cuesta $1 jugar cada juego. Explique su razonamiento.

- (a) Si la proporción de caras es mayor que 0.60, usted gana $1.
- (b) Si la proporción de caras es mayor que 0.40, usted gana $1.
- (c) Si la proporción de caras está entre 0.40 y 0.60, usted gana $1.
- (d) Si la proporción de caras es menor que 0.30, usted gana $1.

3.4 Backgammon. El backgammon es un juego de mesa para dos jugadores en el que las piezas de juego se mueven de acuerdo con el resultado de dos dados. Los jugadores ganan al retirar todas sus piezas del tablero, por lo que suele ser bueno obtener números altos. Está jugando al backgammon con un amigo y saca dos 6 en su primer lanzamiento y dos 6 en su segundo lanzamiento. Su amigo saca dos 3 en su primer lanzamiento y nuevamente en su segundo lanzamiento. Su amigo afirma que está haciendo trampa, porque sacar dobles 6 dos veces seguidas es muy poco probable. Usando la probabilidad, demuestre que sus lanzamientos fueron tan probables como los suyos.

3.5 Lanzamientos de moneda. Si lanza una moneda justa 10 veces, ¿cuál es la probabilidad de

- (a) obtener todas las cruces?
- (b) obtener todas las caras?
- (c) obtener al menos una cruz?

3.6 Lanzamientos de dados. Si lanza un par de dados justos, ¿cuál es la probabilidad de

- (a) obtener una suma de 1?
- (b) obtener una suma de 5?
- (c) obtener una suma de 12?

3.1. DEFINIENDO PROBABILIDAD 93

3.7 Votantes indecisos. Una encuesta de Pew Research preguntó a 2373 votantes registrados seleccionados al azar sobre su afiliación política (Republicano, Demócrata o Independiente) y si se identificaban o no como votantes indecisos. El 35% de los encuestados se identificó como Independiente, el 23% se identificó como votantes indecisos y el 11% se identificó con ambos.21

- (a) ¿Ser Independiente y ser un votante indeciso son eventos disjuntos, es decir, mutuamente excluyentes?
- (b) Dibuje un diagrama de Venn que resuma las variables y sus probabilidades asociadas.
- (c) ¿Qué porcentaje de votantes son Independientes pero no son votantes indecisos?
- (d) ¿Qué porcentaje de votantes son Independientes o votantes indecisos?
- (e) ¿Qué porcentaje de votantes no son ni Independientes ni votantes indecisos?
- (f) ¿El evento de que alguien sea un votante indeciso es independiente del evento de que alguien sea un Independiente político?

3.8 Pobreza e idioma. La Encuesta de la Comunidad Americana es una encuesta continua que proporciona datos cada año para brindar a las comunidades la información actual que necesitan para planificar inversiones y servicios. La Encuesta de la Comunidad Americana de 2010 estima que el 14.6% de los estadounidenses vive por debajo del umbral de la pobreza, el 20.7% habla un idioma que no es inglés (idioma extranjero) en casa y el 4.2% entra en ambas categorías.22

- (a) ¿Vivir por debajo del umbral de la pobreza y hablar un idioma extranjero en casa son eventos disjuntos?
- (b) Dibuje un diagrama de Venn que resuma las variables y sus probabilidades asociadas.
- (c) ¿Qué porcentaje de estadounidenses vive por debajo del umbral de la pobreza y solo habla inglés en casa?
- (d) ¿Qué porcentaje de estadounidenses vive por debajo del umbral de la pobreza o habla un idioma extranjero en casa?
- (e) ¿Qué porcentaje de estadounidenses vive por encima del umbral de la pobreza y solo habla inglés en casa?

- (f) ¿El evento de que alguien viva por debajo del umbral de la pobreza es independiente del evento de que la persona hable un idioma extranjero en casa?

3.9 Disjunto vs. independiente. En las partes (a) y (b), identifique si los eventos son disjuntos, independientes o ninguno (los eventos no pueden ser disjuntos e independientes a la vez).

- (a) Usted y un estudiante seleccionado al azar de su clase obtienen una A en este curso.
- (b) Usted y su compañero de estudio de clase obtienen una A en este curso.
- (c) Si dos eventos pueden ocurrir al mismo tiempo, ¿deben ser dependientes?

3.10 Adivinar en un examen. En un examen de opción múltiple, hay 5 preguntas y 4 opciones para cada pregunta (a, b, c, d). Nancy no ha estudiado nada para el examen y decide adivinar las respuestas al azar. ¿Cuál es la probabilidad de que:

- (a) ¿la primera pregunta que acierta sea la quinta pregunta?
- (b) ¿acierte todas las preguntas?
- (c) ¿acierte al menos una pregunta?

21Pew Research Center, Con los votantes centrados en la economía, la ventaja de Obama se reduce, datos recopilados entre el 4 y el 15 de abril de 2012.

22Oficina del Censo de EE. UU., Estimaciones de 1 año de la Encuesta de la Comunidad Americana de 2010, Características de las personas por idioma Hablado en casa.

		Género	
		Hombre	Mujer
Mayor educación alcanzada	Menos de 9º grado	0.07	0.13
	9º a 12º grado, sin diploma	0.10	0.09
	Graduado de HS (o equivalente)	0.30	0.20
	Algo de universidad, sin título	0.22	0.24
	Título de asociado	0.06	0.08
	Título de licenciatura	0.16	0.17
	Título de posgrado o profesional	0.09	0.09
	Total	1.00	1.00

3.11 Nivel educativo de las parejas. La tabla a continuación muestra la distribución del nivel educativo alcanzado por los residentes de EE. UU. por género, según los datos recopilados en la Encuesta de la Comunidad Americana de 2010.23

- (a) ¿Cuál es la probabilidad de que un hombre elegido al azar tenga al menos una licenciatura?
- (b) ¿Cuál es la probabilidad de que una mujer elegida al azar tenga al menos una licenciatura?

- (c) ¿Cuál es la probabilidad de que un hombre y una mujer que se casan tengan ambos al menos una licenciatura? Tenga en cuenta cualquier suposición que deba hacer para responder a esta pregunta.
- (d) Si hizo una suposición en la parte (c), ¿cree que fue razonable? Si no hizo una suposición, revise su respuesta anterior y luego regrese a esta parte.

3.12 Ausencias escolares. Los datos recopilados en las escuelas primarias del condado de DeKalb, GA, sugieren que cada año aproximadamente el 25% de los estudiantes faltan exactamente un día de escuela, el 15% faltan 2 días y el 28% faltan 3 o más días debido a enfermedad.24

- (a) ¿Cuál es la probabilidad de que un estudiante elegido al azar no falte ningún día de escuela debido a enfermedad este año?
- (b) ¿Cuál es la probabilidad de que un estudiante elegido al azar no falte más de un día?
- (c) ¿Cuál es la probabilidad de que un estudiante elegido al azar falte al menos un día?
- (d) Si un padre tiene dos hijos en una escuela primaria del condado de DeKalb, ¿cuál es la probabilidad de que ninguno de los dos falte a la escuela? Tenga en cuenta cualquier suposición que deba hacer para responder a esta pregunta.
- (e) Si un padre tiene dos hijos en una escuela primaria del condado de DeKalb, ¿cuál es la probabilidad de que ambos niños falten a la escuela, es decir, al menos un día? Tenga en cuenta cualquier suposición que haga.
- (f) Si hizo una suposición en la parte (d) o (e), ¿cree que fue razonable? Si no hizo ninguna suposición, revise sus respuestas anteriores.

23Oficina del Censo de EE. UU., Estimaciones de 1 año de la Encuesta de la Comunidad Americana de 2010, Nivel Educativo.

24S.S. Mizan et al. "Ausencia, Ausencia Extendida y Retraso Repetido Relacionado con el Estado de Asma entre los Niños de Primaria". En: Journal of Asthma 48.3 (2011), pp. 228–234.

3.2 Probabilidad condicional

Puede haber relaciones ricas entre dos o más variables que son útiles para comprender. Por ejemplo, una compañía de seguros de automóviles considerará información sobre el historial de conducción de una persona para evaluar el riesgo de que sea responsable de un accidente. Este tipo de relaciones son el ámbito de las probabilidades condicionales.

3.2.1 Explorando probabilidades con una tabla de contingencia

El conjunto de datos de clasificación de fotos representa un clasificador de una muestra de 1822 fotos de un sitio web para compartir fotos. Los científicos de datos han estado trabajando para mejorar un clasificador para determinar si una foto trata sobre moda o no, y estas 1822 fotos representan una prueba para su clasificador. Cada foto recibe dos clasificaciones: la primera se llama mach learn y proporciona una clasificación de un sistema de aprendizaje automático (ML) de pred fashion o pred not. Cada una de estas 1822 fotos también ha sido clasificada cuidadosamente por un equipo de personas, lo que consideramos la fuente de la verdad; esta variable se llama truth y toma los valores fashion y not. La figura 3.11 resume los resultados.

		truth		
		fashion	not	Total
mach learn	pred fashion	197	22	219
	pred not	112	1491	1603
	Total	309	1513	1822

Figure 3.11: Contingency table summarizing the `photo_classify` data set.

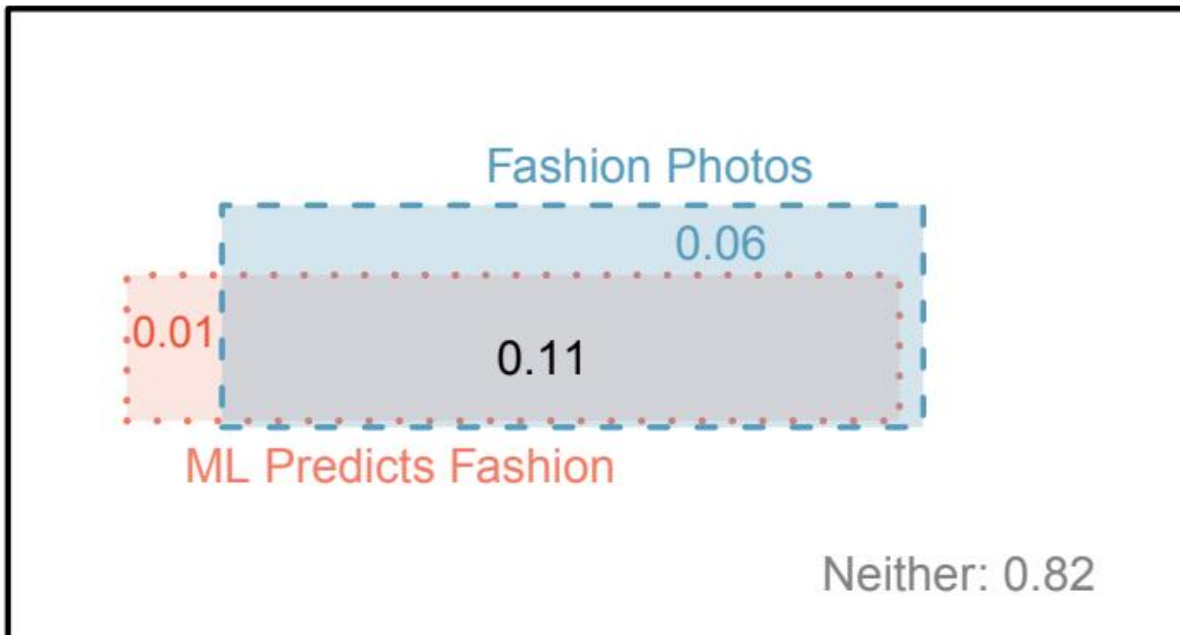


Figura 3.12: Un diagrama de Venn que usa cajas para el conjunto de datos de clasificación de fotos.