1

Introducción a la estadística

PROBLEMA DEL CAPÍTULO

¿Qué podemos aprender de estas encuestas?

A continuación se presentan descripciones breves de cinco encuestas diferentes:

1. A mediados de diciembre de un año reciente, el proveedor de servicios de Internet America Online (AOL) realizó una encuesta entre sus usuarios. La siguiente pregunta se refería a los árboles de Navidad:

   ¿Cuál prefiere?

   - un árbol natural
   - un árbol artificial

   De entre las 7073 respuestas recibidas de los usuarios de Internet, 4650 prefirieron un árbol natural y 2423 un árbol artificial.

2. La revista *Newsweek* hace poco realizó una encuesta acerca del controvertido sitio de Internet llamado Napster, que ofrecía acceso gratuito para copiar discos compactos de música. Se planteó la siguiente pregunta a los lectores:

   ¿Continuaría utilizando Napster si tuviese que pagar?

   Los lectores podían registrar sus respuestas en el sitio de Internet www.newsweek.msnbc.com. De las 1873 respuestas recibidas, el 19% dijo que sí, ya que aun así resultaría más barato que comprar los discos compactos originales. Otro 5% dijo que sí, que se sentiría más cómodo al utilizarlo si lo pagaba.

3. La revista *Good Housekeeping* invitó a mujeres para que visitaran su página en Internet para contestar una encuesta, y se registraron 1500 respuestas. Cuando se les preguntó si preferían tener más dinero o dormir más, el 88% eligió más dinero y sólo el 11%, dormir más.

4. *USA Today* realizó una "Encuesta sobre el cuidado de la salud" de $^3/_4$ de página. A los lectores se les pedía lo siguiente: "Por favor, tómese un momento para llenar esta encuesta y envíenosla". La mayoría de las preguntas se referían a las condiciones de salud, y al consumo de tabaco y de medicamentos de prescripción. La pregunta 17 de la encuesta era: "¿Podríamos establecer contacto nuevamente con usted para que participe en otras encuestas de *USA Today*?".

5. *USA Today* publicó una gráfica de barras con los resultados de una encuesta donde se preguntó a los lectores: "¿Tiene planes para tomar unas vacaciones?". De los 4264 usuarios de Internet que decidieron responder, el 48% dijo que aún no tenía planes, y el 14% contestó que planeaba ir a alguna playa.

¿Qué característica importante tienen en común estas cinco encuestas? Con base en los resultados obtenidos en ellas, ¿cómo se ven afectadas nuestras conclusiones respecto de la población general? ¿Podríamos concluir que la mayoría de los estadounidenses prefieren un árbol de Navidad real que uno artificial? ¿Concluiríamos que la gran mayoría de las mujeres estadounidenses prefieren más dinero que dormir más? ¿O que la gran mayoría de las mujeres lectoras de la revista *Good Housekeeping* prefieren más dinero que dormir más? Las respuestas a tales preguntas son de crucial importancia para evaluar los resultados de las encuestas. El asunto a considerar aquí es el tema más importante de todo este capítulo y podría ser el aspecto más relevante de todo el libro.

En este capítulo estudiaremos temas relevantes sobre la validez de encuestas como las anteriormente descritas. Veremos que con frecuencia sacamos conclusiones relevantes con la simple aplicación del sentido común. Al final de este capítulo, seremos capaces de identificar los aspectos clave que afectan la validez de las cinco encuestas y lograremos una profunda comprensión de los métodos de recolección de datos en general.

1-1 Panorama general

El Problema del capítulo en la página anterior implica a las encuestas. La encuesta es una de muchas herramientas disponibles para recolectar datos. Una meta común de las encuestas es reunir datos de una pequeña parte de un grupo más grande para aprender algo acerca de este último. Una meta común e importante de la estadística es aprender acerca de un grupo examinando los datos de algunos de sus miembros. En dicho contexto los términos *muestra* y *población* adquieren importancia. Las definiciones formales de éstos y otros términos básicos se presentan a continuación.

Definiciones

Datos son las observaciones recolectadas (como mediciones, géneros, respuestas de encuesta).

Estadística es una colección de métodos para planear experimentos, obtener datos, y después organizar, resumir, presentar, analizar, interpretar y llegar a conclusiones basadas en los datos.

Población es la colección completa de todos los elementos (puntuaciones, personas, mediciones, etcétera) a estudiar. Se dice que la colección es completa, pues incluye a todos los sujetos que se estudiarán.

Censo es la colección de datos de *cada uno* de los miembros de la población.

Muestra es un *subconjunto* de miembros seleccionados de una población.

Por ejemplo, un sondeo de Gallup preguntó a 1087 adultos: “¿Consume bebidas alcohólicas como licor, vino o cerveza o es abstemio?”. Los 1087 sujetos de la encuesta constituyen una *muestra* mientras que la *población* consiste en el conjunto de los 202,682,345 estadounidenses adultos. Cada 10 años el gobierno de Estados Unidos intenta obtener un *censo* de cada ciudadano; pero no logra hacerlo porque es imposible localizar a cada uno de ellos. En la actualidad hay polémica en torno al intento de emplear métodos estadísticos acertados para aumentar la exactitud del censo, aunque los aspectos políticos constituyen un factor clave para que los miembros del Congreso se resistan a esta mejoría. Quizás algún día algunos lectores de este texto sean miembros del Congreso y tengan la sabiduría de traer el censo al siglo XXI.

Una función importante de este libro es demostrar cómo utilizar las muestras de datos para llegar a conclusiones respecto de poblaciones. Veremos que es *extremadamente* importante obtener datos muestrales que sean representativos de la población de la que se tomaron. Por ejemplo, si usted encuesta a los estudiantes graduados de su universidad y les pide que anoten sus ingresos anuales y le envíen la respuesta por correo, es probable que los resultados no sean representativos de la población de todo el alumnado. Aquellos con bajos ingresos estarían menos inclinados a responder y quienes respondan pueden mostrar tendencia a exagerar.

Al avanzar en este capítulo debemos enfocarnos en los siguientes conceptos clave:

- **Los datos muestrales deben reunirse de una forma adecuada, como en un proceso de selección *aleatoria*.**
- **Si los datos muestrales no se reúnen de forma adecuada, resultarían tan inútiles que ninguna cantidad de tortura estadística podría salvarlos.**

Ante todo, le pedimos que comience a estudiar estadística con una mente abierta. No considere que el estudio de la estadística es comparable con un procedimiento inflexible. La experiencia del autor es que los estudiantes a menudo se sorprenden por lo interesante que resulta la estadística y también porque realmente llegan a dominar sus principios básicos sin mucha dificultad, incluso si no han sido sobresalientes en otros cursos de matemáticas. Estamos convencidos de que cuando usted termine este curso introductorio, tendrá la firme creencia de que la estadística es una materia rica e interesante con aplicaciones que son extensivas, reales y significativas. También estamos convencidos de que con la asistencia a clases y la dedicación constantes, usted tendrá éxito para dominar los conceptos básicos de la estadística presentados en este curso.

1-2 Tipos de datos

En la sección 1-1 definimos los términos *población* y *muestra*. Los siguientes dos términos se utilizan para distinguir los casos donde se cuenta con los datos de una población completa, de aquellos en que sólo se tienen datos de una muestra.

Definiciones

Parámetro es una medición numérica que describe algunas características de una *población*.

Estadístico es una medición numérica que describe algunas características de una *muestra*.

EJEMPLOS

1. **Parámetro:** Cuando Lincoln fue elegido presidente por primera vez, recibió el 39.82% de 1,865,908 votos. Si suponemos que el conjunto de todos esos votos es la población a considerar, entonces el 39.82% es un *parámetro*, no un estadístico.
2. **Estadístico:** Con base en una muestra de 877 ejecutivos encuestados, se encontró que el 45% de ellos no contrataría a alguien con un error ortográfico en su solicitud de empleo. Esta cifra del 45% es un *estadístico*, ya que está basada en una muestra, no en la población completa de todos los ejecutivos.

El estado de la estadística

El término *estadística* se deriva de la palabra latina *status* (que significa "estado"). Los primeros usos de la estadística implicaron la compilación de datos y la elaboración de gráficas para describir diversos aspectos de un estado o de un país. En 1662, John Graunt publicó información estadística acerca de los nacimientos y los decesos. Al trabajo de Graunt siguieron estudios de tasas de mortalidad y de enfermedad, tamaño de poblaciones, ingresos y tasas de desempleo. Los hogares, gobiernos y negocios se apoyan bastante en datos estadísticos para dirigir sus acciones. Por ejemplo, se compilan datos cuidadosamente y con regularidad para establecer las tasas de desempleo, las tasas de inflación, los índices del consumidor y las tasas de nacimiento y muerte, y los líderes empresariales utilizan los datos resultantes para tomar decisiones que afectan las futuras contrataciones, los niveles de producción y la expansión hacia nuevos mercados.

Algunos conjuntos de datos consisten en números (como estaturas de 66 y 72 pulgadas), mientras que otros son no numéricos (como los colores de ojos verde y café). Los términos *datos cuantitativos* y *datos cualitativos* suelen utilizarse para distinguir entre ambos tipos.

Definiciones

Los **datos cuantitativos** consisten en números que representan conteos o mediciones.

Los **datos cualitativos** (o **categóricos** o **de atributo**) se dividen en diferentes categorías que se distinguen por alguna característica no numérica.

EJEMPLOS

1. **Datos cuantitativos:** Los pesos de las supermodelos.
2. **Datos cualitativos:** El género (hombre/mujer) de atletas profesionales.

Cuando se trabaja con datos cuantitativos, es importante utilizar las unidades de medida apropiadas, tales como dólares, horas, pies, metros y otras. Debemos ser especialmente cuidadosos para observar aquellas referencias como "todas las cantidades están en *miles de dólares*" o "todos los tiempos están en *centésimas de segundo*" o "las unidades están en *kilogramos*". Ignorar unidades de medida como éstas podría llevar a conclusiones incorrectas. La NASA perdió su Mars Climate Orbiter de 125 millones de dólares cuando la sonda se estrelló, porque la programación de control tenía los datos de aceleración en unidades *inglesas*, pero ellos incorrectamente consideraron que estaban en unidades *métricas*.

Los datos cuantitativos se describen con mayor detalle distinguiendo entre los tipos *discretos* y *continuos*.

Definiciones

Datos discretos resultan cuando el número de posibles valores es un número finito, o bien, un número que puede contarse. (Es decir, el número de posibles valores es 0, 1, 2, etcétera).

Datos continuos (numéricos) resultan de un infinito de posibles valores que pueden asociarse a puntos de alguna escala continua, cubriendo un rango de valores sin huecos ni interrupciones.

EJEMPLOS

1. **Datos discretos:** Las cantidades de huevos que ponen las gallinas son datos *discretos* porque representan conteos.

2. **Datos continuos:** Las cantidades de leche que las vacas producen son datos *continuos* porque son mediciones que pueden tomar cualquier valor dentro de un intervalo continuo. Durante un intervalo de tiempo dado, una vaca producirá una cantidad de leche que puede ser cualquier valor entre 0 y 5 galones. Es posible obtener 2.343115 galones, ya que la vaca no está restringida a producir cantidades discretas de 0, 1, 2, 3, 4, o 5 galones.

Otra forma común de clasificación de los datos es el uso de cuatro niveles de medición: nominal, ordinal, de intervalo y de razón. Cuando la estadística se aplica a problemas reales, el nivel de medición de los datos es un factor importante para determinar el procedimiento a usar. (Véase la figura 14.1 en la página 727.) En este libro encontraremos algunas referencias a estos niveles de medición; sin embargo, lo importante aquí es sustentarse en el sentido común: no hay que hacer cálculos ni usar métodos estadísticos con datos que no sean apropiados. Por ejemplo, no tendría sentido calcular un promedio de números del seguro social, ya que estos números son datos que se usan como identificación, y no representan mediciones ni conteos de algo. Por la misma razón, no tendría sentido calcular un promedio de los números que aparecen en las camisetas de los jugadores de básquetbol.

Definición

Nivel de medición nominal son los datos consistentes exclusivamente en nombres, etiquetas o categorías que no pueden acomodarse según un esquema de orden (por ejemplo, de bajo a alto).

EJEMPLOS Los ejemplos siguientes ilustran datos muestrales en el nivel de medición nominal.

1. **Sí/no/indeciso:** Respuestas de sí, no e indeciso en una encuesta.
2. **Colores:** Los colores de automóviles conducidos por estudiantes universitarios (rojo, negro, azul, blanco y otros).

Puesto que los datos nominales carecen de un orden o de un significado numérico, no pueden utilizarse para realizar cálculos. A veces se asignan números a las diferentes categorías (en especial cuando los datos se codifican para el uso de sistemas de cómputo), pero tales números no tienen significado computacional y cualquier promedio que se calcule carece de sentido.

Medición de la desobediencia

¿De qué manera se recolectan datos que parecen imposibles de medir, como el nivel de desobediencia de las personas? El psicólogo Stanley Milgram ideó el siguiente experimento: un investigador enseñó a un sujeto voluntario a operar un tablero de control que administraba "choques eléctricos" cada vez más dolorosos a una tercera persona. En realidad no se daban tales choques y la tercera persona era un actor. El voluntario iniciaba con 15 volts y fue instruido para incrementar los choques en aumentos de 15 volts. El nivel de desobediencia fue el punto donde el sujeto se negaba a incrementar el voltaje. Resultó sorprendente que dos terceras partes de los sujetos obedecieran las órdenes aun cuando el actor gritaba y fingía sufrir un ataque cardiaco.

Definición

Los datos están en el **nivel de medición ordinal** cuando pueden acomodarse en algún orden, aunque no es posible determinar diferencias entre los valores de los datos o tales diferencias carecen de significado.

EJEMPLOS Los siguientes son ejemplos de datos muestrales en el nivel de medición ordinal.

1. **Las calificaciones de un curso:** Un profesor universitario asigna calificaciones de A, B, C, D, o E, las cuales pueden acomodarse en orden; sin embargo, no es posible determinar diferencias entre ellas. Por ejemplo, sabemos que A es más alto que B (por lo tanto, existe un orden), pero no podemos restar B de A (por lo tanto, no se calcula la diferencia).
2. **Rangos ordenados:** Con fundamento en varios criterios, una revista clasificó las ciudades de acuerdo con su "calidad de vida"**.** Tales rangos (primero, segundo, tercero, etcétera) determinan un orden; sin embargo, las diferencias entre los rangos ordenados no tienen significado alguno. Por ejemplo, una diferencia de "segundo menos primero" puede sugerir 2 – 1 = 1, pero este resultado de 1 no tiene significado porque no es una cantidad exacta que pueda compararse con otras diferencias del mismo tipo. La diferencia entre la primera ciudad y la segunda no es la misma que la diferencia entre la segunda y la tercera. Utilizando los rangos ordenados de la revista, la *diferencia* entre las ciudades de Nueva York y Boston no puede compararse cuantitativamente con la *diferencia* entre las ciudades de San Luis y Filadelfia.

Los datos ordinales ofrecen información sobre comparaciones relativas, aunque no sobre las magnitudes de las diferencias. Por lo general, los datos ordinales no se usan para cálculos como un promedio, pero esta norma se quebranta en ocasiones (como cuando se usan calificaciones con letras para calcular el punto promedio de calificación).

Apuesta por la ciencia

En ocasiones los datos se recolectan de maneras muy ingeniosas y de fuentes muy extrañas. Un ejemplo es el de ciertos investigadores que estudiaron los cambios climáticos. Ellos se dieron cuenta de que cada primavera, desde 1917, en la pequeña ciudad de Nenana, Alaska, hacían un juego de lotería, en el cual las personas apostaban sobre la hora exacta en que la capa de hielo del río Tanana se rompería (el último premio fue de cerca de 300,000 dólares). Se colocó un tripié en el río congelado y éste se conectó a un reloj. El reloj se detendría cuando el hielo, al quebrarse, moviera el tripié. De esta forma los investigadores supieron el momento preciso en que ocurría la rotura cada año desde 1917, y los datos resultaron muy útiles en el estudio de las tendencias climáticas.

Definición

El **nivel de medición de intervalo** se parece al nivel ordinal, pero con la propiedad adicional de que la diferencia entre dos valores de datos cualesquiera tiene un significado. Sin embargo, los datos en este nivel no tienen un punto de partida inherente (*natural)* desde cero (donde *nada* de la cantidad esté presente).

EJEMPLOS Los siguientes ejemplos ilustran el nivel de medición de intervalo.

1. **Temperaturas:** Las temperaturas corporales de 98.2°F y 98.6°F son ejemplos de datos en este nivel de medición. Tales valores están ordenados, y podemos determinar su diferencia de 0.4°F. Sin embargo, no existe un punto de partida natural. El valor de 0°F quizá parezca un punto de partida, pero es arbitrario y no representa la ausencia total de calor. Como 0°F no es

un punto de partida desde cero natural, es erróneo decir que 50°F es *dos veces* más caliente que 25°F.

2. **Años:** Los años 1000, 2000, 1776 y 1492. (El tiempo no inició en el año 0, así que el año 0 es arbitrario en vez de ser un punto de partida de cero natural, que representaría "ausencia de tiempo").

Definición

El **nivel de medición de razón** se parece al nivel de intervalo, aunque tiene la propiedad adicional de que sí tiene un punto de partida o cero inherente (donde cero indica que *nada* de la cantidad está presente). Para valores en este nivel, tanto las diferencias como las proporciones tienen significado.

EJEMPLOS Los siguientes son ejemplos de datos en el nivel de medición de razón. Observe la presencia del valor cero natural y el uso de proporciones que significan "dos veces" y "tres veces".

1. **Pesos:** Los pesos (en quilates) de anillos engastados con diamante (0 efectivamente representa ausencia de peso y 4 quilates es dos veces el peso de 2 quilates).
2. **Precios:** Los precios de los libros de texto universitarios ($0 efectivamente representa ningún costo y un libro de $90 es tres veces más costoso que un libro de $30).

Este nivel de medición se denomina "de razón" porque el punto de partida cero hace que las razones o cocientes tengan significado. Entre los cuatro niveles de medición, la mayoría de las dificultades surgen con la distinción entre los niveles de intervalo y de razón.

Sugerencia: Para hacer más fácil esta distinción, utilice una sencilla "prueba de razón": considere dos cantidades en las cuales un número es dos veces el otro y pregúntese si "dos veces" se puede usar para describir correctamente las cantidades. Puesto que un peso de 200 libras es *dos veces* más pesado que un peso de 100 libras, pero 50°F *no es dos veces* más caliente que 25°F, los pesos están en el nivel de razón, mientras que las temperaturas Fahrenheit están en el nivel de intervalo. Para una comparación y un repaso concisos, estudie la tabla 1-1 en la página siguiente, que señala las diferencias entre los cuatro niveles de medición.

1-2 Destrezas y conceptos básicos

En los ejercicios 1 a 4, determine si el valor dado es un estadístico o un parámetro.

1. El Senado actual de Estados Unidos consta de 87 hombres y 13 mujeres.
2. Se selecciona una muestra de estudiantes y el número promedio (media) de libros de texto comprados este semestre es 4.2.
3. Se toma una muestra de estudiantes y el promedio (media) de la cantidad de tiempo de espera en la fila para comprar libros de texto este semestre es 0.65 horas.
4. En un estudio de los 2223 pasajeros del *Titanic*, se encontró que 706 sobrevivieron cuando se hundió.

Tabla 1-1 Niveles de medición de datos

Nivel	Resumen	Ejemplo	
Nominal	Sólo rangos de orden. Los datos no pueden acomodarse en un esquema de orden.	Origen de estudiantes: 5 californianos 20 texanos 40 neoyorquinos	Sólo rangos de orden o nombres.
Ordinal	Rangos de orden que pueden acomodarse, pero no hay diferencias o carecen de significado.	Automóviles de estudiantes: 5 compactos 20 medianos 40 grandes	Orden determinado por "compacto, mediano, grande".
De intervalo	Las diferencias son significativas, pero no hay punto de partida natural y las razones no tienen significado.	Temperaturas del campus: 5°F 20°F 40°F	0°F no es "sin calor". 40°F no es dos veces más caliente que 20°F.
De razón	Hay un punto de partida natural y las razones tienen significado.	Distancias de viaje de estudiantes: 5 km 20 km 40 km	40 km es *dos veces* más lejos que 20 km.

En los ejercicios 5 a 8, determine si los valores dados provienen de un conjunto de datos discreto o continuo.

5. El salario presidencial de George Washington era de 25,000 dólares anuales y el salario presidencial actual es de 400,000 anuales.

6. Un estudiante de estadística obtiene datos muestrales y encuentra que la media del peso de automóviles en la muestra es 3126 libras.

7. En una encuesta de 1059 adultos, se encontró que el 39% de ellos tienen pistolas en sus casas (de acuerdo con una encuesta de Gallup).

8. Cuando se probaron 19,218 máscaras antigás de divisiones de la milicia de Estados Unidos, se encontró que 10,322 estaban defectuosas (de acuerdo con datos de la revista *Time*).

En los ejercicios 9 a 16, determine cuál de los cuatro niveles de medición (nominal, ordinal, de intervalo, de razón) es el más apropiado.

9. Las estaturas de las mujeres que juegan básquetbol en la WNBA.

10. Las calificaciones de fantástico, bueno, promedio, pobre o inaceptable en citas a ciegas.

11. Las temperaturas actuales en los salones de clase en su universidad.

12. Los números en las camisetas de las mujeres que juegan básquetbol en la WNBA.

13. Las calificaciones de la revista *Consumer Reports* de "mejor compra, recomendado, no recomendado".

14. Los números del seguro social.

15. El número de respuestas "sí" recibidas cuando se les preguntó a 1250 conductores si habían usado alguna vez un teléfono celular mientras conducían.

16. Los códigos postales de la ciudad en que vive.

En los ejercicios 17 a 20, identifique a) *la muestra y* b) *la población. También determine si la muestra parece ser representativa de la población.*

17. Un reportero de *Newsweek* se para en una esquina y pregunta a 10 adultos si creen que el presidente actual está haciendo un buen trabajo.

18. Nielsen Media Research encuesta a 5000 amas de casa seleccionadas al azar y encuentra que el 19% de los televisores encendidos están sintonizados en *60 minutos* (de acuerdo con datos de *USA Today*).

19. En una encuesta de Gallup aplicada a 1059 adultos seleccionados aleatoriamente, el 39% respondió "sí" cuando se le preguntó: "¿Tiene usted una pistola en su casa?".

20. Una estudiante graduada de la Universidad de Newport realizó un proyecto de investigación acerca de cómo se comunican los adultos estadounidenses. Empezó por una encuesta que envió por correo a 500 de los adultos que ella conocía. Les pidió que le enviaran por correo la respuesta a esta pregunta: "¿Prefiere usted usar el correo electrónico o el correo tortuga (el servicio postal estadounidense)?". Ella recibió a vuelta de correo 65 respuestas y 42 de ellas indicaron una preferencia por el correo tortuga.

1-2 Más allá de lo básico

21. Interpretación de los incrementos de temperatura En la tira cómica "Born Loser" de Art Sansom, Brutus se alegra por un incremento en la temperatura de 1° a 2°. Cuando alguien le pregunta qué tiene de bueno estar a 2°, él responde que "hace dos veces más calor que en la mañana". Explique por qué Brutus está equivocado una vez más.

22. Interpretación de encuesta política Un encuestador aplica una encuesta a 200 personas y les pregunta por el partido político de su preferencia: él codifica las respuestas como 0 (para demócrata), 1 (para republicano), 2 (para independiente) y 3 (para otras respuestas cualesquiera). Entonces calcula el promedio (media) de los números y obtiene 0.95. ¿Cómo se interpreta este valor?

23. Escala para calificar comida Un grupo de estudiantes desarrolló una escala para calificar la calidad de la comida de la cafetería de su escuela, donde 0 representaba "neutral: ni buena ni mala". Se asignaron números negativos a las comidas malas y números positivos a las comidas buenas; la magnitud del número correspondía a la severidad de lo bueno o lo malo. Las primeras tres comidas se calificaron con 2, 4 y -5. ¿Cuál es el nivel de medición de calificaciones como éstas? Explique su respuesta.

1-3 Pensamiento crítico

El éxito en el curso introductorio de estadística por lo regular requiere de más *sentido común* que destreza matemática (a pesar de la advertencia de Voltaire de que "el sentido común no es muy común"). Ya que ahora tenemos acceso a calculadoras y a computadoras, las aplicaciones modernas de la estadística ya no requieren

¿Debe creerse en un estudio estadístico?

En la segunda edición del libro *Statistical Reasoning for Everyday Life*, los autores Jeff Bennet, William Briggs y Mario Triola enumeran las siguientes directrices para evaluar de forma crítica un estudio estadístico: **1.** Identifique la meta del estudio, la población considerada y el tipo de estudio. **2.** Considere la fuente, particularmente respecto de la posibilidad de la existencia de prejuicios. **3.** Analice el método de obtención de muestras. **4.** Busque problemas en la definición o medición de variables de interés. **5.** Tenga cuidado con variables confusas que podrían invalidar las conclusiones. **6.** Considere el escenario y la redacción de cualquier encuesta. **7.** Verifique que las gráficas representen los datos con fidelidad y que las conclusiones tengan justificación. **8.** Considere si las conclusiones logran los objetivos del estudio, si tienen sentido y si tienen un significado práctico.

que dominemos algoritmos complejos de operaciones matemáticas. En su lugar, nos enfocamos en la *interpretación* de los datos y los resultados. Esta sección está diseñada para ilustrar la forma en que se usa el sentido común cuando pensamos de forma crítica acerca de los datos y la estadística.

Hace cerca de un siglo, el estadista Benjamin Disraeli pronunció la famosa frase: "Hay tres clases de mentiras: mentiras, viles mentiras y estadísticas". También se ha dicho que "las cifras no mienten; los mentirosos calculan las cifras". El historiador Andrew Lang dijo que algunas personas utilizan la estadística "como un borracho utiliza los postes de alumbrado: como apoyo más que como iluminación". El caricaturista político Don Wright nos anima diciendo "retome el misterio de la vida: mienta a un encuestador". El autor Franklin P. Jones escribió que "la estadística puede usarse para sustentar cualquier cosa, en especial a los estadísticos". En el *Esar's Comic Dictionary* encontramos la definición de que un estadístico es "un especialista que reúne pensamientos y luego los conduce al extravío". Estas afirmaciones se refieren a ejemplos donde los métodos estadísticos se utilizaron de forma errónea, de manera que resultaron engañosos en última instancia. Hay dos fuentes principales de tal engaño: **1.** el intento malintencionado por parte de personas deshonestas, y **2.** los errores de descuido cometidos por personas que no conocen nada mejor. Sin tener en cuenta la fuente, como ciudadanos responsables y como empleados profesionales valiosos, debemos tener una habilidad básica para distinguir entre conclusiones estadísticas que parecen ser válidas de las que son gravemente defectuosas.

Para mantener esta sección en la perspectiva apropiada, hay que saber que éste no es un libro acerca de los malos usos de la estadística. El resto de este libro estará lleno de usos muy importantes de métodos estadísticos válidos. Aprenderemos métodos generales para usar datos muestrales y así poder hacer inferencias relevantes acerca de poblaciones; aprenderemos acerca de encuestas y tamaños de muestra, acerca de mediciones importantes de características fundamentales de los datos. Junto con las explicaciones de estos conceptos generales, veremos muchas aplicaciones específicas reales, tales como los efectos en el fumador pasivo, el predominio del alcohol y el tabaco en las películas de dibujos animados para niños y la calidad de productos de consumo, incluyendo dulces M&M, cereales, Coca Cola y Pepsi. Pero incluso en estas aplicaciones reales y con significado, debemos ser cuidadosos para interpretar correctamente los resultados de métodos estadísticos válidos.

Comenzamos nuestro desarrollo del pensamiento crítico considerando muestras erróneas. Estas muestras son erróneas en el sentido de que el método de muestreo arruina la muestra, de modo que tiene la posibilidad de estar sesgada (es decir, de no ser representativa de la población de la que se obtuvo). La sección siguiente analiza con más detalle los métodos de muestreo y describe la importancia de la *aleatoriedad*. El primer ejemplo sigue un procedimiento de muestreo que tiene una seria carencia de aleatoriedad, la cual es muy importante. La siguiente definición se refiere a uno de los usos incorrectos de la estadística más comunes y graves.

Definición

Muestra de respuesta voluntaria (o **muestra autoseleccionada**) es aquella donde los sujetos deciden ser incluidos por sí mismos.

1 Para ver ejemplos, remítase al problema del capítulo. Cuando America Online o alguien más aplica una encuesta por Internet, los individuos por sí mismos deciden si participan o no, así que constituyen una muestra de respuesta voluntaria. Sin embargo,

existen mayores posibilidades de que las personas con opiniones decididas participen, de tal forma que las respuestas no sean representativas de toda la población. A continuación se presentan ejemplos de muestras de respuesta voluntaria que, por su naturaleza, adolecen de una carencia importante, pues no debemos obtener conclusiones sobre una población con base en una muestra sesgada como ésta:

- Las encuestas manejadas a través de Internet, en las que los sujetos deciden si responden o no.
- Las encuestas por correo, donde los sujetos deciden si contestan.
- Las encuestas telefónicas, en las que anuncios en el periódico, la radio, o la televisión, le piden que tome un teléfono voluntariamente y llame a un número especial para registrar su opinión.

Con muestras de respuesta voluntaria como éstas, sólo es posible llegar a conclusiones válidas acerca del grupo específico que decide participar; pero sería una práctica incorrecta común establecer conclusiones acerca de una población más grande. Desde un punto de vista estadístico, una muestra como ésta falla en lo esencial y no debe utilizarse para realizar declaraciones generales acerca de una población mayor.

Muestras pequeñas Las conclusiones no deben basarse en muestras que son sumamente pequeñas. Por ejemplo, el Children's Defense Fund publicó *Children Out of School in America*, donde se reportó que de los estudiantes de escuela secundaria suspendidos en una región, el 67% fueron suspendidos al menos tres veces. ¡Pero esta cifra está basada en una muestra de sólo *tres* estudiantes! Los reportes en los medios de comunicación fallaron al mencionar que el tamaño de la muestra era muy pequeño. (En los capítulos 6 y 7 veremos que *en ocasiones* es posible realizar algunas deducciones valiosas a partir de muestras pequeñas, aunque debemos ser cuidadosos y verificar que se satisfagan los requisitos necesarios).

En ocasiones una muestra puede parecer relativamente grande (como en una encuesta de "2000 adultos estadounidenses seleccionados al azar"), pero si se obtienen conclusiones acerca de los subgrupos, por ejemplo, los republicanos de sexo masculino de 21 años de edad de Pocatello, tales conclusiones estarían basadas en muestras demasiado pequeñas. Si bien es importante tener una muestra que sea suficientemente grande, también lo es el hecho de tener datos muestrales que se recolecten de una forma adecuada, como la selección aleatoria. Aun las muestras grandes llegan a ser muestras erróneas.

Gráficas Las gráficas —como las de barras y las circulares— en ocasiones sirven para exagerar o disfrazar la verdadera naturaleza de los datos. (En el capítulo 2 analizaremos una variedad de gráficas diferentes). Las dos gráficas en la figura 1-1 de la siguiente página representan *los mismos datos* del Bureau of Labor Statistics, aunque el inciso *b*) está diseñado para exagerar la diferencia entre los salarios semanales de hombres y mujeres. Al no iniciar el eje vertical en cero, la gráfica del inciso *b*) tiende a producir una impresión subjetiva engañosa, que hace que los lectores incorrectamente crean que la diferencia es mucho peor de lo que en realidad es. La figura 1-1 enseña una lección importante: para interpretar una gráfica de manera correcta, debemos analizar la información *numérica* dada en ella, para no engañarnos por su forma general. (El término mediana que se utiliza en la figura 1-1 se describirá con claridad en la sección 2-4).

Pictogramas Los dibujos de objetos, llamados pictogramas, también pueden resultar engañosos. Algunos objetos que se usan comúnmente para representar datos incluyen objetos tridimensionales, como bolsas de dinero, pilas de monedas, tanques militares (para gastos militares), barriles (para producción petrolera) y casas

La encuesta del Literary Digest

En la contienda presidencial de 1936, la revista *Literary Digest* efectuó una encuesta y predijo la victoria de Alf Landon, pero Franklin D. Roosevelt obtuvo una victoria abrumadora. Maurice Bryson señala: "Se enviaron 10 millones de papeletas de muestra para votar a prospectos de votantes, aunque sólo se devolvieron 2.3 millones. Como todo el mundo debía saber, tales muestras prácticamente siempre están sesgadas". Bryson también afirma: "La respuesta voluntaria a cuestionarios enviados por correo es tal vez el método más común que los estadísticos han encontrado para recolectar datos en las ciencias sociales, y tal vez sea también el peor". (Véase el artículo de Bryson "The *Literary Digest* Poll: Making of a Statistical Myth", *The American Statistician*, vol. 30, núm. 4).

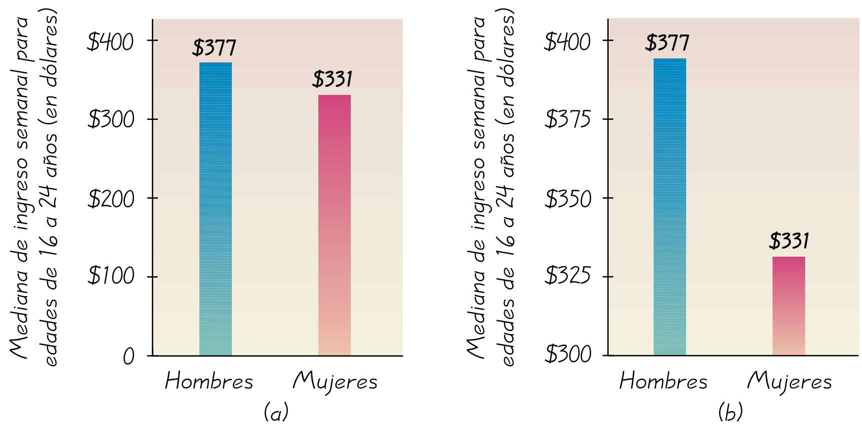


FIGURA 1-1 Salarios semanales de hombres y mujeres de 16 a 24 años

FIGURA 1-2 Pictograma

Duplique el largo, el ancho y la altura de un cubo y el volumen se incrementa por un factor de ocho, como se indica. Si el cubo más pequeño representa los impuestos en un año y el cubo más grande representa el *doble* de los impuestos algún tiempo después, los últimos impuestos parecen ser ocho veces más grande, y no dos, la cantidad original.

(para construcción de viviendas). Al dibujar tales objetos, los artistas llegan a crear impresiones falsas que distorsionan las diferencias. Si duplicamos cada lado de un cuadrado, el área no tan sólo se duplica, sino que aumenta en un factor de cuatro. Si se duplica cada lado de un cubo, el volumen no se duplica simplemente, sino que se incrementa en un factor de ocho, como se observa en la figura 1-2. Si los impuestos se duplican durante una década, un artista podría representar las cantidades de impuestos con una bolsa de dinero para el primer año y otra bolsa de dinero dos veces más ancha, dos veces más alta y dos veces más profunda para el segundo año. En vez de parecer que los impuestos se duplican, parecerá que aumentaron en un factor de ocho y así el dibujo distorsionaría la verdad.

Porcentajes A veces se utilizan porcentajes engañosos o poco claros. Si usted toma el 100% de alguna cantidad, está tomándolo *todo*. (No debería requerir de un 110% de esfuerzo para que la declaración anterior tenga sentido). En referencia a la pérdida de equipaje, la Continental Airlines publicó anuncios afirmando que se trata de "un área en la que ya hemos mejorado un 100% en los últimos seis meses". En un editorial que criticaba ese dato estadístico, el diario *The New York Times* interpretó correctamente que la cifra de mejora en un 100% significa que ya no se está perdiendo equipaje, logro que todavía no disfruta Continental Airlines.

Los siguientes son algunos principios clave que se aplican cuando tratamos con porcentajes. Todos estos principios usan la noción básica de que % o "por ciento" significa realmente "dividido entre 100". Este primer principio se empleará con frecuencia en este libro.

- **Porcentaje de:** Para encontrar el *porcentaje de* una cantidad, excluya el símbolo % y divida el valor del porcentaje entre 100, después multiplique por la cantidad. Este ejemplo muestra que el 6% de 1200 es 72:

$$\text{el 6\% de 1200 respuestas} = \frac{6}{100} \times 1200 = 72$$

- **Fracción → Porcentaje:** *Para convertir de una fracción a un porcentaje*, divida el denominador entre el numerador para obtener un número decimal equivalente, y después multiplíquelo por 100 y agregue el símbolo %. Este ejemplo muestra que la fracción es equivalente al 75%:

$$\frac{3}{4} = 0.75 \rightarrow 0.75 \times 100\% = 75\%$$

- **Decimal → Porcentaje:** *Para convertir de un número decimal a un porcentaje*, multiplíquelo por 100%. Este ejemplo demuestra que 0.234 es equivalente a 23.4%:

$$0.234 \rightarrow 0.234 \times 100\% = 23.4\%$$

- **Porcentaje → Decimal:** *Para convertir de un porcentaje a un número decimal*, elimine el símbolo % y divida entre 100. Este ejemplo demuestra que 85% es equivalente a 0.85:

$$85\% = \frac{85}{100} = 0.85$$

Preguntas predispuestas Existen muchos aspectos que afectan las preguntas de una encuesta. Éstas llegan a estar "cargadas" o redactadas intencionalmente de manera que propicien una respuesta deseada. Observe las calificaciones de respuesta "sí" reales para las diferentes redacciones en una pregunta:

- 97% sí: "¿Debe el presidente utilizar su poder de veto para eliminar los desperdicios?".
- 57% sí: "¿Debe el presidente utilizar su poder de veto o no?"

En *The Superpollsters*, David W. Moore describe un experimento donde se preguntó a diferentes sujetos si estaban de acuerdo con las siguientes declaraciones:

- Se gasta muy poco dinero en subsidios del Estado.
- Se gasta muy poco dinero en asistencia a los pobres.

Aun cuando es el pobre quien recibe el subsidio del Estado, sólo el 19% estuvo de acuerdo cuando se usaron las palabras "subsidio del Estado", aunque el 63% estuvo de acuerdo con "asistencia a los pobres".

Orden de las preguntas En ocasiones las preguntas de una encuesta se cargan de forma no intencional, en virtud de factores como el orden de los reactivos que se someten a consideración. Observe estas preguntas de una encuesta aplicada en Alemania:

- ¿Cree usted que el tránsito vehicular contribuye a la contaminación del aire más o menos que la industria?
- ¿Cree usted que la industria contribuye a la contaminación del aire más o menos que el tránsito vehicular?

Cuando se presentó primero el tránsito, el 45% culpó al tránsito y el 27% culpó a la industria; cuando la industria se presentó primero, el 24% culpó al tránsito y el 57% culpó a la industria.

Rechazo Cuando se invita a las personas a contestar una encuesta, algunas se niegan con firmeza a responder. La tasa de rechazo ha crecido en años recientes, en parte porque muchos vendedores persistentes de empresas de telemercadeo buscan vender bienes o servicios comenzando con una inducción de ventas que suena como si fuera parte de una encuesta de opinión. En *Lies, Damn Lies, and Statistics*, el autor Michael Wheeler indica con acierto que "las personas que se niegan a hablar con los entrevistadores parecen ser diferentes de quienes no lo hacen. Algunas quizá tengan miedo a los extraños y otras sean celosas de su privacidad, pero su negativa a hablar demuestra que su visión del mundo circundante es marcadamente diferente de aquellas otras personas que permiten a los entrevistadores entrar a sus hogares".

LA ESTADÍSTICA EN LAS NOTICIAS

La estadística y las minas terrestres

La International Campaign to Ban Land Mines (la Campaña Internacional para Proscribir Minas Terrestres) y el director ejecutivo de la Vietnam Veterans of America Foundation (VVAF) fueron recientemente galardonados con el Premio Nobel de la Paz. Cuando la VVAF pidió ayuda en la recolección de datos acerca de las minas terrestres, se reunió a un equipo de notables estadísticos. En vez de trabajar con datos intangibles, como el valor de la vida humana, ellos trabajaron con datos tangibles en bruto, como es el área que inutiliza un campo minado y el costo de cultivos que no se cosechan. Los datos se incluyeron en el libro *After the Guns Fall Silent: The Enduring Legacy of Landmines*, que vino a ser un recurso clave en las discusiones del tema de las minas terrestres. El *AMSTAT News* citó a uno de los editores del libro: "Este esfuerzo de reunión y análisis de datos es lo que hizo posible presentar el tema ante los legisladores. El trabajo en verdad marcó una diferencia".

Detección de datos falsos

Un maestro de clase asigna la tarea de registrar los resultados de lanzar al aire una moneda 500 veces. Un estudiante deshonesto decide ahorrar tiempo inventando los resultados, en lugar de realmente lanzar la moneda. Como las personas generalmente no pueden inventar resultados que en realidad sean aleatorios, con frecuencia identificamos datos falsos como éstos. En 500 lanzamientos de una moneda real, es en extremo probable que usted obtenga una serie de seis caras o seis cruces, aunque la gente casi nunca incluye una racha como ésta cuando inventa resultados.

Otra forma de detectar datos “fabricados” consiste en establecer que los resultados violan la ley de Benford: para muchos grupos de datos, los primeros dígitos no están uniformemente distribuidos. Más bien los primeros dígitos de 1, 2, . . . , 9 ocurren con frecuencia de 30%, 18%, 12%, 10%, 8%, 7%, 6%, 5% y 5%, respectivamente. (Véase “The Difficulty of Faking Data” por Theodore Hill, *Chance*, vol. 12, núm. 3).

Correlación y causalidad En el capítulo 9 de este libro analizaremos la asociación estadística entre dos variables, como son la riqueza y el CI. Usaremos el término correlación para indicar que las dos variables están relacionadas. Sin embargo, en el capítulo 9 hacemos esta importante anotación: *la correlación no implica causalidad*. Esto significa que cuando nosotros encontramos una asociación estadística entre dos variables, no podemos concluir que una de las variables es la causa de la otra (o que la afecta directamente). Si encontramos una correlación entre la riqueza y el CI, no podemos concluir que el CI de una persona afecta directamente su riqueza, ni tampoco podemos concluir que la riqueza de la persona afecta directamente su puntuación de CI. En los medios de comunicación es bastante común reportar una correlación recién encontrada con una redacción que indica o implica directamente que una de las variables es causa de la otra.

Estudios para el propio beneficio Algunas veces los estudios reciben el patrocinio de grupos con intereses específicos que buscan promover. Por ejemplo, Kiwi Brands, un fabricante de abrillantador de calzado, encargó un estudio que suscitó esta declaración impresa en algunos periódicos: “De acuerdo con una encuesta nacional realizada a 250 empleadores profesionales, la razón más común del fracaso de un solicitante de trabajo del sexo masculino al dar una buena primera impresión, fue llevar los zapatos desaseados”. Debemos ser muy cautos con encuestas como éstas, cuyos resultados generan ganancias económicas para el patrocinador. En los últimos años ha generado preocupación creciente la práctica de las compañías farmacéuticas de financiar a doctores que realizan experimentos clínicos y reportan sus resultados en revistas de prestigio, como *Journal of American Medical Association.*

Números precisos “En la actualidad existen 103,215,027 hogares en Estados Unidos.” Puesto que esta cantidad es muy precisa, mucha gente considera erróneamente que también es exacta. En este caso, ese número es un estimado y sería mejor decir que el número de hogares es de alrededor de 103 millones.

Imágenes parciales “El 90% de todos nuestros automóviles, vendidos en este país en los últimos 10 años, continúa circulando”. Millones de consumidores escucharon ese anuncio comercial y no se dieron cuenta de que el 90% de los automóviles que el anunciante vendió en este país se vendieron durante los últimos tres años, de modo que la mayoría de esos automóviles que circulaban estaban casi nuevos. La afirmación era técnicamente correcta, aunque muy engañosa, al no presentar los resultados completos.

Distorsiones deliberadas En el libro *Tainted Truth*, Cynthia Crossen cita un ejemplo de la revista *Corporate Travel* que publicó resultados que mostraban que, entre las compañías de renta de automóviles, Avis fue la ganadora en una encuesta realizada a personas que utilizan ese servicio. Cuando Hertz solicitó información detallada acerca de la encuesta, las respuestas originales de ésta desaparecieron y el coordinador de encuestas de la revista renunció. Hertz demandó a Avis (por publicidad falsa basada en la encuesta) y a la revista; al final las compañías llegaron a un acuerdo.

Además de los casos ya citados, se conocen muchos otros usos incorrectos de la estadística; algunos de estos otros casos se encuentran en libros como el clásico de Darrel Huff, *How to Lie with Statistics*; el de Robert Reichard, *The Figure Finaglers*, y el de Cynthia Crossen, *Tainted Truth*. Comprender tales prácticas resultará extremadamente útil en la evaluación de los datos estadísticos que se encuentran en situaciones cotidianas.

1-3 Destrezas y conceptos básicos

En los ejercicios 1 a 4, utilice el pensamiento crítico para desarrollar una conclusión alternativa. Por ejemplo, considere un reporte de medios de comunicación donde los conductores de BMW gozan de mejor salud que los adultos que no manejan. La conclusión de que los automóviles BMW son la causa de una mejor salud quizás esté equivocada. La siguiente sería una mejor conclusión: Los conductores de BMW tienden a ser más adinerados que los adultos que no manejan y una mayor riqueza se asocia con una mejor atención a la salud.

1. El peso y los camiones Un estudio demostró que los conductores de camiones pesan más que los adultos que no manejan camiones. Conclusión: Los camiones causan que la gente gane peso.

2. Las casas y la longevidad Un estudio concluyó que los propietarios de casas tienden a vivir más tiempo que quienes no habitan viviendas propias. Conclusión: Poseer una casa crea paz y armonía internas que causan que las personas tengan mejor estado de salud y vivan más tiempo.

3. Cumplimiento de las leyes de tránsito Un estudio mostró que en el condado de Orange se expidieron más multas por exceso de velocidad a personas de grupos minoritarios que a los blancos. Conclusión: En el condado de Orange las personas de grupos minoritarios conducen a mayor velocidad que los blancos.

4. Remedio para el resfriado En un estudio de síntomas del resfriado, se encontró que cada uno de los sujetos de estudio con resfriado mejoró dos semanas después de tomar píldoras de jengibre. Conclusión: las píldoras de jengibre curan el resfriado.

En los ejercicios 5 a 16, utilice el pensamiento crítico para señalar el tema principal.

5. El chocolate es un alimento saludable El *New York Times* publicó un artículo que incluyó esta afirmación: "Por fin, el chocolate ocupa el lugar que merece en la pirámide de los alimentos, junto a sus vecinos de clase alta, el vino tinto, las frutas, los vegetales y el té verde. Varios estudios reportados en el *Journal of Nutrition* mostraron que, después de comer chocolate, los sujetos a prueba incrementaron los niveles de antioxidantes en su sangre. El chocolate contiene flavonoides, antioxidantes asociados con la disminución del riesgo de enfermedades cardiacas y derrame cerebral. Mars Inc., la compañía de dulces, y la Chocolate Manufacturers Association financiaron gran parte de la investigación". ¿Qué está equivocado en este estudio?

6. Datos de censo Después de realizado el último censo nacional, el *Poughkeepsie Journal* imprimió este titular de primera página: "281,421,906 en Estados Unidos". ¿Qué está mal en este titular?

7. Encuesta por correo Cuando la autora Shere Hite escribió *Woman and Love: A Cultural Revolution in Progress*, basó sus conclusiones en 4500 respuestas recibidas después de enviar por correo 100,000 cuestionarios a varios grupos de mujeres. ¿Es probable que sus conclusiones sean válidas, en el sentido de que puedan aplicarse a la población general de todas las mujeres? ¿Por qué sí o por qué no?

8. Números "900" En una encuesta de *Nightline* de la ABC, 186,000 televidentes pagaron 50 centavos cada uno para llamar a un número telefónico "900" y dar su opinión acerca de mantener la sede de las Naciones Unidas en Estados Unidos. Los resultados demostraron que el 67% de quienes llamaron estuvieron a favor de que las Naciones Unidas salieran de Estados Unidos. Interprete los resultados identificando lo que concluiríamos acerca del sentir de la población general, respecto de mantener la sede de las Naciones Unidas en Estados Unidos.

9. Encuestas telefónicas La Hartford Insurance Company lo contrató a usted para encuestar a una muestra de adultos acerca de sus compras de automóviles. ¿Cuál es el error al considerar a las personas cuyos números telefónicos aparecen listados en los directorios como población de la cual se toma la muestra?

10. Crimen y autobuses El *Newport Chronicle* afirma que los paraderos de autobús causan crímenes, porque un estudio concluyó que las tasas de crimen son más altas en las ciudades con paraderos de autobús, que en las zonas rurales que carecen de ellos. ¿Cuál es el error en esta afirmación?

11. Cascos de motocicleta El Senado del estado de Hawai entró en audiencia para considerar una ley que obligaba a los motociclistas a usar cascos. Algunos motociclistas testificaron que habían participado en choques donde los cascos habían resultado inútiles. ¿Qué grupo importante no fue capaz de testificar? (Véase "A Selection of Selection Anomalies" de Wainer, Palmer y Bradlow en *Chance*, vol. 11, núm. 2).

12. La encuesta al cliente de Merrill Lynch El autor recibió una encuesta de la empresa de inversiones Merrill Lynch. La encuesta fue diseñada para medir su satisfacción como cliente y contenía preguntas específicas para calificar al consultor financiero personal del autor. La portada de la carta incluyó esta declaración: "Sus respuestas son extremadamente valiosas para su consultor financiero, Russell R. Smith, y para Merrill Lynch.... Compartiremos su nombre y las respuestas con su consultor financiero". ¿Cuál es el error en esta encuesta?

13. La nicotina de los cigarrillos Remítase al conjunto de datos 5 del Apéndice B y considere el contenido de nicotina de 29 diferentes marcas de cigarrillos. El promedio (media) de esas cantidades es 0.94 mg. ¿Es probable que este resultado sea un buen estimado del promedio (media) de todos los cigarrillos que se han fumado en Estados Unidos? ¿Por qué sí o por qué no?

14. Pregunta incorrecta Una encuesta incluye este reactivo: "Anote su altura en pulgadas". A partir de este dato se espera obtener las estaturas reales de los encuestados y analizarlas, aunque hay dos problemas básicos diferentes en este reactivo; identifíquelos.

15. Longevidad Usted necesita realizar un estudio de longevidad a personas que nacieron después del fin de la Segunda Guerra Mundial en 1945. Si usted visitara los cementerios y utilizara las fechas de nacimiento y muerte indicadas en las lápidas, ¿obtendría buenos resultados? ¿Por qué sí o por qué no?

16. SMSI En una carta al editor del *New York Times*, la ciudadana de Moorestown, New Jersey, Jean Mercer criticó la declaración de que "colocar a los bebés en posición supina ha disminuido las muertes por SMSI". SMSI son las siglas del Síndrome de Muerte Súbita Infantil y la posición supina implica estar tendido sobre la espalda con la cara hacia arriba. Ella sugirió que esta afirmación es mejor: "Los pediatras aconsejaron la posición supina durante un periodo en que disminuyeron las tasas de SMSI". ¿Qué está equivocado al decir que la posición supina ha disminuido las muertes por SMSI?

En los ejercicios 17 a 22, conteste las preguntas que se hacen en relación con los porcentajes.

17. Porcentajes
a. Convierta la fracción 17/25 a un porcentaje equivalente.
b. Convierta 35.2% a su equivalente decimal.
c. ¿Cuánto es el 57% de 1500?
d. Convierta 0.486 a un porcentaje equivalente.

18. Porcentajes
a. ¿Cuál es el 26% de 950?
b. Convierta 5% en su equivalente decimal.
c. Convierta 0.01 a un porcentaje equivalente.
d. Convierta la fracción 527/1200 a un porcentaje equivalente. Redondee la respuesta a la décima más cercana del porcentaje.

19. Porcentajes en una encuesta de Gallup

a. En una encuesta de Gallup, el 52% de 1038 adultos entrevistados dijo que ser un fumador pasivo es "muy dañino". ¿Cuál es el número real de adultos que dicen que ser un fumador pasivo es "muy dañino"?

b. De entre los 1038 adultos entrevistados, 52 dijeron que ser un fumador pasivo "no es dañino en absoluto". ¿Cuál es el porcentaje de gente que escogió "no es dañino en absoluto"?

20. Porcentajes en un estudio del Lipitor

a. En un estudio del fármaco Lipitor para el colesterol, a 270 pacientes se les administró un placebo; 19 de esos 270 pacientes reportaron dolor de cabeza. ¿Qué porcentaje de este grupo placebo reportó dolor de cabeza?

b. De entre los 270 pacientes del grupo placebo, el 3.0% reportó dolores de espalda. ¿Cuál es el número real de pacientes que reportaron dolores de espalda?

21. Porcentajes delictivos en los planteles universitarios Un estudio de los delitos cometidos por estudiantes bajo la influencia del alcohol o las drogas en los planteles universitarios, se basó en una encuesta por correo a 1875 estudiantes. Un artículo del *USA Today* destacó que "el 8% de los estudiantes, que respondieron de manera anónima, afirmaron haber cometido un delito en el campus. Y el 62% de ese grupo dijo que lo hizo bajo la influencia del alcohol o las drogas". Considerando que el número de estudiantes que respondió de manera anónima es 1875, ¿cuántos cometieron realmente un delito en el campus mientras estaban bajo la influencia del alcohol o las drogas?

22. Porcentajes en los medios de comunicación

a. Un editorial del *New York Times* criticó un gráfico que describía un enjuague bucal que "reduce la placa bacteriana en más del 300%". ¿Que es incorrecto en esta declaración?

b. En el *New York Times Magazine*, un reporte acerca de la disminución de la inversión occidental en Kenia afirmó que "después de años de vuelos diarios, Lufthansa y Air France han interrumpido el servicio de pasajeros. La inversión extranjera cayó el 500% durante la década de 1990". ¿Qué está equivocado en esta declaración?

1-3 Más allá de lo básico

23. Datos falsos Un investigador del Sloan-Kettering Cancer Research Center fue criticado por falsificar datos. Entre sus datos había cifras obtenidas de seis grupos de ratones, con 20 ratones en cada grupo. Estos valores se dieron para el porcentaje de éxito en cada grupo: 53%, 58%, 63%, 46%, 48%, 67%. ¿Cuál es la principal falla?

24. ¿Qué está mal en el asunto? Trate de identificar cada una de las cuatro fallas principales en lo siguiente. Un diario realizó una encuesta pidiendo a los lectores que llamaran y respondieran esta pregunta: "¿Apoya usted el desarrollo de armas atómicas que podrían matar a millones de personas inocentes?". Se reportó que 20 lectores respondieron y 87% contestó "no", mientras que el 13% dijo "sí".

25. Redacción predispuesta Escriba una pregunta de encuesta que trate sobre un tema de su interés. Primero redacte la pregunta con objetividad, después redáctela para fomentar las respuestas hacia cierta dirección y por tercera vez redáctela para influir en las respuestas en la dirección opuesta.

26. Gráficas Actualmente, las mujeres ganan 74 centavos por cada dólar que ganan los hombres al realizar el mismo trabajo. Dibuje una gráfica que describa esta información de manera objetiva; luego, dibuje una gráfica que exagere la diferencia. (*Sugerencia:* Consulte la figura 1-1.)

1-4 Diseño de experimentos

Si bien esta sección contiene mucha información, existen dos puntos principales que son bastante sencillos. Es necesario entender que el método usado para reunir los datos es extremadamente importante, y debemos reconocer que la *aleatoriedad* resulta importante en particular.

- **Si los datos muestrales no se reúnen de manera adecuada, éstos podrían resultar inútiles por completo, de tal forma que ninguna cantidad de tortura estadística los salvaría.**
- **La *aleatoriedad* por lo general juega un papel crucial para determinar cuáles son los datos a reunir.**

Los métodos estadísticos se rigen por los datos. Por lo regular obtenemos datos de dos fuentes distintas: los *estudios observacionales* y los *experimentos*.

Definiciones

En un **estudio observacional**, observamos y medimos características específicas, aunque no intentamos *manipular* a los sujetos que estamos estudiando.

En un **experimento** aplicamos algún *tratamiento* y luego procedemos a observar sus efectos sobre los sujetos.

Una encuesta de Gallup es un buen ejemplo de un estudio observacional, mientras que la prueba clínica del fármaco Lipitor es un buen ejemplo de un experimento. La encuesta de Gallup es observacional en el sentido de que simplemente se observan personas (a menudo por medio de entrevistas) sin modificarlas de ninguna forma. Pero la prueba clínica de Lipitor implica el tratamiento de algunas personas con el fármaco, de manera que se manipula a los sujetos tratados. Hay diferentes tipos de estudios observacionales, como se ilustra en la figura 1-3. Estos términos, que se usan comúnmente en muchas y diferentes revistas profesionales, se definen aquí.

Definiciones

En un **estudio transeccional**, los datos se observan, miden y reúnen en un solo momento.

En un **estudio retrospectivo** (o **de control de caso**), los datos se toman del pasado (a través del examen de registros, entrevistas y otros medios).

En un **estudio prospectivo** (o **longitudinal** o **cohorte**), los datos se reúnen en el futuro y se toman de grupos (llamados *cohortes*) que comparten factores comunes.

Existe una distinción importante entre el muestreo realizado en estudios retrospectivos y estudios prospectivos. En los estudios retrospectivos regresamos en el tiempo a reunir datos acerca de características resultantes que nos conciernen, como un grupo de conductores que murieron en accidentes automovilísticos y otro grupo de conductores que no murieron en este tipo de accidentes. En los estudios

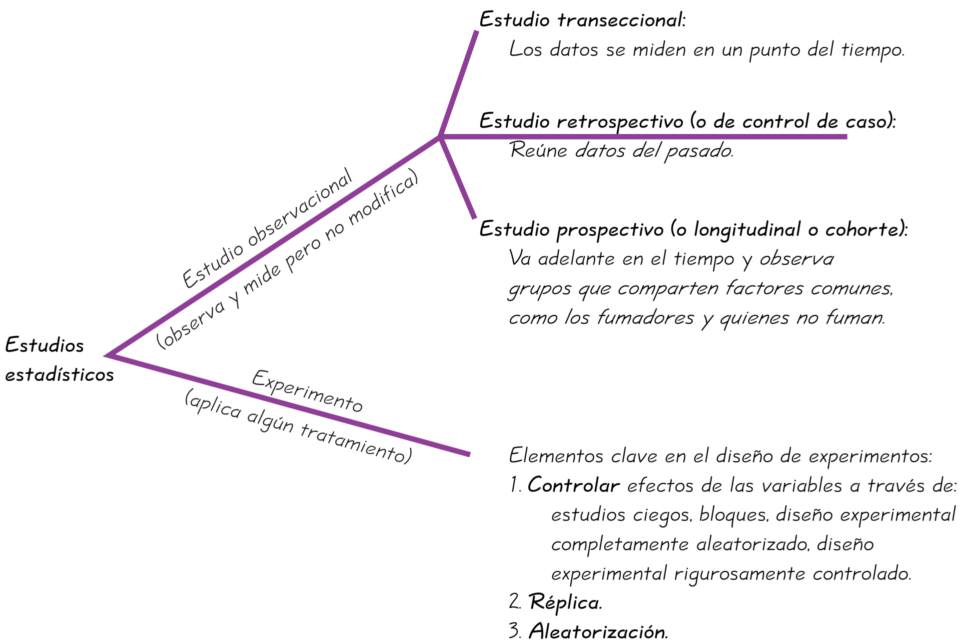


FIGURA 1-3 Elementos de los estudios estadísticos

prospectivos vamos adelante en el tiempo siguiendo grupos con un factor causal potencial y grupos que no lo tienen, como un grupo de conductores que utilizan teléfonos celulares y un grupo de conductores que no usan teléfonos celulares.

Las tres definiciones se aplican a los estudios observacionales, aunque por ahora nos enfocaremos en los experimentos. Los resultados de los experimentos algunas veces se empobrecen a causa de la *confusión*.

Definición

La **confusión** ocurre en un experimento cuando el experimentador no es capaz de distinguir entre los efectos de diferentes factores.

Intente planear el experimento de manera que no se presente confusión.

Por ejemplo, suponga que un profesor de Vermont experimenta con una nueva política de asistencia ("su calificación promedio en el curso bajará un punto por cada clase que falte"); sin embargo, llega un invierno excepcionalmente benigno que carece de nieve y temperaturas muy frías, lo cual en años anteriores obstaculizó la asistencia. Si la asistencia mejora no será posible determinar si la mejoría es atribuible a la nueva política de asistencia o al invierno benigno. Se confunden los efectos de la política de asistencia y del clima.

Control de los efectos de las variables

La figura 1-3 muestra que uno de los elementos clave en el diseño de experimentos es controlar los efectos de las variables. Se adquiere este control utilizando dispositivos como el estudio ciego, los bloques, el diseño experimental completamente aleatorizado o un diseño experimental rigurosamente controlado, que se describen a continuación.

LA ESTADÍSTICA EN LAS NOTICIAS

Pruebas clínicas *vs.* estudios observacionales

En un artículo del *New York Times* acerca de la terapia hormonal para las mujeres, la reportera Denise Grady escribió acerca de un reporte de tratamientos probados en ensayos controlados aleatorizados. Ella declaró que "pruebas como ésta, donde los pacientes se designan al azar para un tratamiento o un placebo, se consideran el estándar por excelencia en la investigación médica. En contraste, los estudios observacionales, en los que los pacientes deciden por sí mismos si toman un fármaco, se consideran menos confiables... Los investigadores manifiestan que los estudios observacionales tal vez han dado una falsa imagen color de rosa del reemplazo hormonal, ya que las mujeres que optan por recibir los tratamientos son más saludables y tienen mejores hábitos al empezarlos que las mujeres que no lo hacen".

Los efectos Hawthorne y del experimentador

El conocido efecto placebo ocurre cuando un sujeto no tratado cree incorrectamente que está recibiendo un tratamiento real e informa una mejoría en sus síntomas. El efecto Hawthorne ocurre cuando, por alguna razón, los sujetos tratados responden de manera diferente por el simple hecho de ser parte del experimento. (Este fenómeno se denominó "efecto Hawthorne" porque se observó por primera vez en un estudio realizado con obreros en la planta Hawthorne, de Western Electric). Ocurre un efecto del experimentador (a veces llamado efecto Rosenthall) cuando el investigador o experimentador influye, sin desearlo, en los sujetos mediante factores como la expresión facial, el tono de voz o la actitud.

Estudio ciego En 1954 se diseñó un experimento masivo para probar la efectividad de la vacuna de Salk en la prevención de la poliomielitis que mató o paralizó a miles de niños. En este experimento a un grupo de tratamiento se le administró la vacuna real de Salk, mientras a un segundo grupo se le dio un placebo que no contenía ningún fármaco. En los experimentos que involucran placebos, hay a menudo un **efecto placebo** que ocurre cuando un sujeto no tratado reporta una mejoría en los síntomas. (La mejoría reportada en el grupo placebo puede ser real o imaginaria). Este efecto placebo llega a minimizarse o a tomarse en cuenta mediante el uso del **estudio ciego**, una técnica donde el sujeto no sabe si está recibiendo un tratamiento o un placebo. El estudio ciego nos permite determinar si el efecto del tratamiento es significativamente diferente del efecto placebo. El experimento de la poliomielitis fue un **estudio doble ciego**, lo que quiere decir que el estudio ciego ocurrió a dos niveles: **1.** los niños inyectados no sabían si estaban recibiendo la vacuna de Salk o un placebo, y **2.** los doctores que suministraron las inyecciones y evaluaron los resultados tampoco lo sabían.

Bloques Cuando se diseña un experimento para probar la efectividad de uno más tratamientos, es importante poner a los sujetos (con frecuencia llamados *unidades experimentales*) en diferentes grupos (o *bloques*), de manera que estos grupos sean muy similares. Un **bloque** es un grupo de sujetos que son similares en formas que podrían afectar el resultado del experimento.

Cuando realice un experimento con el objetivo de probar uno o más tratamientos diferentes, forme bloques (o grupos) de sujetos con características similares.

Diseño experimental completamente aleatorizado Cuando se decide cómo asignar a los sujetos a los diferentes bloques, se puede utilizar una selección aleatoria o intentar controlar cuidadosamente la asignación, para que los sujetos de cada bloque resulten similares. Una opción consiste en usar un **diseño experimental completamente aleatorizado,** mediante el cual los sujetos se asignan a los diferentes bloques a través de un proceso de *selección aleatoria*. Un ejemplo de un diseño experimental completamente aleatorizado es el experimento de la poliomielitis: los niños fueron asignados al grupo de tratamiento o al grupo placebo a través de un proceso de selección aleatoria (equivalente a lanzar una moneda al aire).

Diseño rigurosamente controlado Otra opción para asignar sujetos a los bloques es el uso del **diseño rigurosamente controlado**, donde los sujetos son *cuidadosamente elegidos* para que quienes formen cada bloque sean similares en las características que sean importantes para el experimento. En un experimento para probar la efectividad de un fármaco para disminuir la presión sanguínea, si el grupo placebo incluye a una persona del sexo masculino de 30 años de edad, con sobrepeso, fumador, con alto consumo de bebidas alcohólicas y con una dieta alta en sal y grasas, el grupo de tratamiento también debe incluir a una persona con características similares (lo cual, en este caso, sería fácil de conseguir).

Réplica y tamaño de muestra

Además de controlar los efectos de las variables, otro elemento clave del diseño experimental es el tamaño de las muestras. Éstas deben ser suficientemente grandes para que el comportamiento errático, que es característico de muestras muy pequeñas, no disfrace los efectos verdaderos de los diferentes tratamientos. La repetición de un experimento se llama **réplica,** la cual se utiliza con efectividad cuando tenemos los sujetos suficientes como para reconocer las diferencias que resultan de los diferentes tratamientos. (En otro contexto, la *réplica* se refiere a la repetición o duplicación de un experimento para confirmar o verificar los resultados). Con la réplica se incrementa la posibilidad de reconocer diferentes efectos del tratamiento en los tamaños de muestra grandes. Sin embargo, una muestra grande no es necesariamente una muestra buena. Aunque es importante tener una muestra que sea suficientemente grande, es más importante tener una muestra en la que los datos se escojan de una forma apropiada, como la selección aleatoria (que se describirá después).

Utilice un tamaño de muestra que sea lo bastante grande para distinguir la verdadera naturaleza de cualquiera de los diferentes efectos, y obtenga la muestra usando un método adecuado, como uno basado en la *aleatoriedad*.

En el experimento diseñado para probar la vacuna de Salk, a 200,000 niños se les administró la vacuna de Salk real, y a otros 200,000 niños se les dio un placebo. Se observó la efectividad de la vacuna porque se usaron tamaños de muestra bastante grandes en el experimento real. No obstante, aunque los grupos de tratamiento y placebo fueran muy grandes, el experimento puede fallar si los sujetos no se asignan a los dos grupos de tal manera que ambos grupos sean similares en las características importantes para el experimento.

Aleatorización y otras estrategias de muestreo

En la estadística, como en la vida, uno de los peores errores es reunir datos en una forma que no sea la adecuada. Insistiremos en este punto muy importante:

Si los datos muestrales no se reúnen de forma adecuada, resultarían tan inútiles que ninguna cantidad de tortura estadística podrá salvarlos.

En la sección 1-3 vimos que una muestra de respuesta voluntaria es aquella donde los sujetos deciden por sí mismos si responden o no. Este tipo de muestras son muy comunes, aunque sus resultados por lo general resultan inútiles para hacer inferencias válidas acerca de poblaciones más grandes.

Ahora definiremos algunos de los métodos de muestreo más comunes.

Definiciones

En una **muestra aleatoria** los miembros de una población se seleccionan de manera que cada *miembro individual* tiene la misma posibilidad de ser elegido.

Una **muestra aleatoria simple** del tamaño de n sujetos, se selecciona de manera que cada posible *muestra del mismo tamaño n* tenga la misma posibilidad de ser elegida.

EJEMPLO Muestra aleatoria y muestra aleatoria simple Imagine un salón de clases con 60 estudiantes acomodados en seis filas de 10 estudiantes cada una. Suponga que el profesor selecciona una muestra de 10 estudiantes tirando un dado y seleccionando la fila correspondiente al resultado. ¿El resultado es una muestra aleatoria? ¿Es una muestra aleatoria simple?

SOLUCIÓN La muestra es una muestra aleatoria porque cada estudiante tiene la misma posibilidad (una posibilidad en seis) de ser elegido. Sin embargo, la muestra no es una muestra aleatoria simple porque no todas las muestras de tamaño 10 tienen la misma posibilidad de ser escogidas. Por ejemplo, este diseño muestral de usar un dado para seleccionar una fila hace imposible seleccionar 10 estudiantes que estén en filas diferentes (aunque hay una posibilidad en seis de seleccionar la muestra que consiste en los 10 estudiantes en la primera fila).

Importante: A lo largo de este libro utilizaremos una variedad de procedimientos estadísticos diferentes y muchas veces tendremos como requisito reunir una *muestra aleatoria simple*, como se define arriba.

Con el muestreo aleatorio se espera que todos los componentes de la población estén (aproximadamente) representados de manera proporcional. Las muestras aleatorias se seleccionan mediante diversos métodos, incluyendo el uso de computadoras para generar números aleatorios. (Antes del uso de las computadoras, las tablas de números aleatorios se utilizaban con frecuencia. Si quiere leer algo verdaderamente interesante, consulte el libro *A million random digits*, publicado por Free Press, que contiene un millón de dígitos generados aleatoriamente. El resumen del argumento no está disponible todavía). A diferencia de un muestreo realizado con descuido o por casualidad, el muestreo aleatorio exige una muy cuidadosa planeación y ejecución.

Además del muestreo aleatorio, hay otras técnicas de muestreo en uso, y las más comunes se describen aquí. Observe la figura 1-4, una ilustración que describe los diferentes tipos de muestreo. Tome en cuenta que sólo el muestreo aleatorio y el muestreo aleatorio simple se usarán en el resto de este libro.

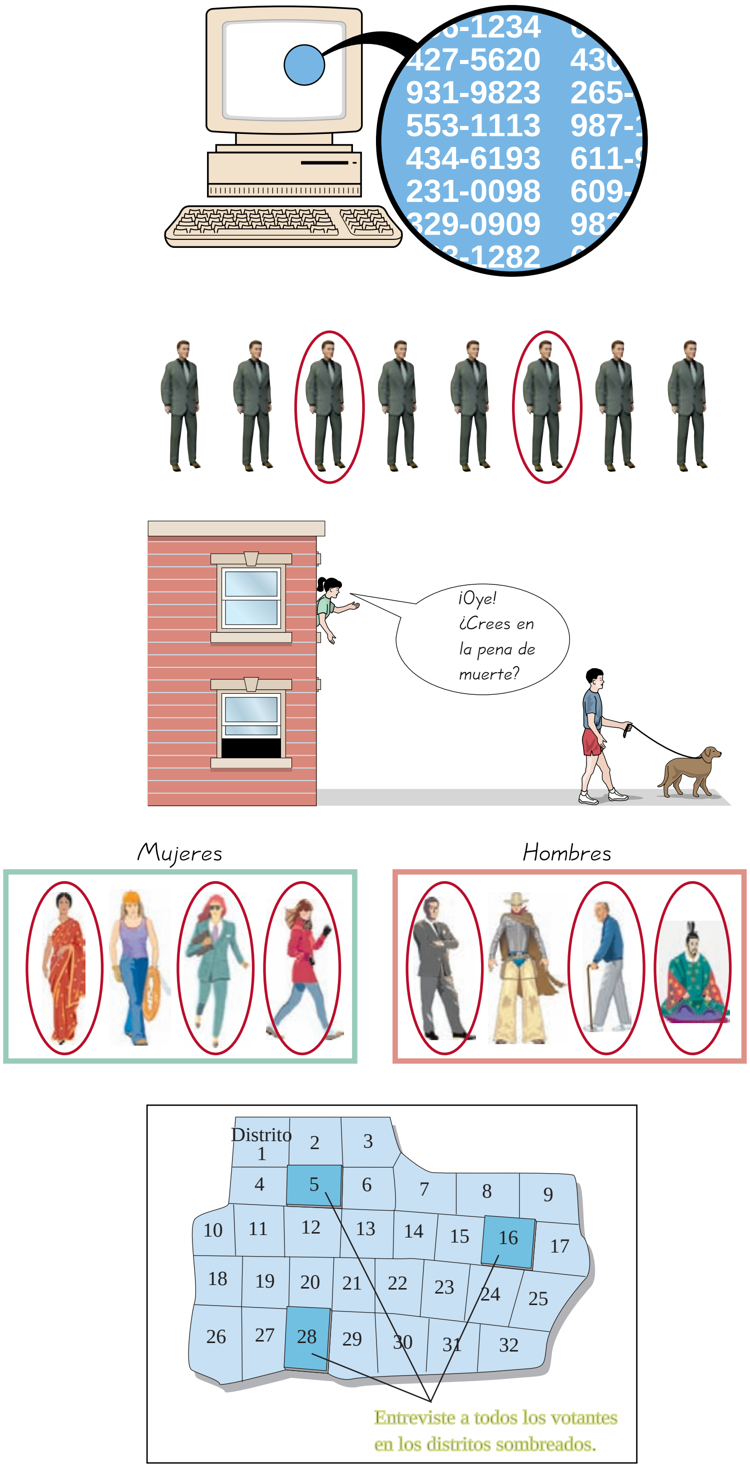
Definiciones

En el **muestreo sistemático**, elegimos algún punto de partida y luego seleccionamos cada *k*-ésimo (por ejemplo cada quincuagésimo) elemento en la población.

Con el **muestreo de conveniencia**, simplemente se utilizan resultados que sean muy fáciles de obtener.

Con el **muestreo estratificado**, subdividimos la población en al menos dos diferentes subgrupos (o estratos) que comparten las mismas características (por ejemplo, el género o la categoría de edad) y después realizamos un muestreo de cada subgrupo (o estrato).

En el **muestreo por racimos**, primero dividimos el área de la población en secciones (o racimos), después seleccionamos aleatoriamente algunos de estos racimos, y luego elegimos a *todos* los miembros de los racimos seleccionados.



Muestreo aleatorio:
Cada miembro de la población tiene la misma probabilidad de ser seleccionado. A menudo se usan computadoras para generar números telefónicos aleatorios.

Muestreo aleatorio simple:
Se selecciona una muestra de tamaño de n sujetos de manera que cada posible muestra del mismo tamaño n tenga la misma posibilidad de ser elegida.

Muestreo sistemático:
Se selecciona un punto de partida, después se selecciona cada k-ésimo (por ejemplo, cada quincuagésimo) elemento en la población.

Muestreo de conveniencia:
Se utilizan resultados que son fáciles de obtener.

Muestreo estratificado:
Se subdivide a la población en al menos dos diferentes subgrupos (o estratos) que comparten las mismas características (por ejemplo, el género o categoría de edad), y después se extrae una muestra de cada subgrupo.

Muestreo por racimos:
Se divide el área de la población en secciones (o racimos), se eligen al azar unas cuantas de estas secciones y luego se escogen todos los miembros de los racimos seleccionados.

FIGURA 1-4 **Métodos de muestreo comunes**

Es fácil confundir el muestreo estratificado y el muestreo por racimos, ya que ambos suponen la formación de subgrupos. Pero el muestreo por racimos usa *todos* los miembros de una *muestra* de racimos, mientras el muestreo estratificado usa una *muestra* de los miembros de *todos* los estratos. Un ejemplo de muestreo por racimos es una encuesta previa a las elecciones, donde se seleccionan aleatoriamente 30 distritos electorales de un número mayor de distritos, luego se encuesta a todas las personas de cada uno de esos distritos escogidos, lo cual es mucho más rápido y mucho menos costoso que seleccionar a una persona de cada uno de los muchos distritos del área de población. Los resultados de la muestra estratificada o por racimos se ajustan o se ponderan para corregir cualquier representación desproporcionada de los grupos.

Para un tamaño de muestra fijo, si usted selecciona sujetos de diferentes estratos al azar, es probable que obtenga resultados más consistentes (y menos variables) que si simplemente selecciona una muestra al azar de la población general. Por esta razón, el muestreo estratificado se utiliza con frecuencia para reducir la variación en los resultados. Muchos de los métodos que se analizarán después en este libro tienen como requisito que los datos muestrales constituyen una *muestra aleatoria simple*, y ni el muestreo estratificado ni el muestreo por racimos satisfacen este requisito.

La figura 1-4 ilustra métodos de muestreo comunes. Los profesionales a veces reúnen datos usando cierta combinación de tales métodos. Aquí está un ejemplo típico de lo que se llama un *diseño muestral de etapas múltiples*: primero se selecciona una muestra aleatoria de condados de todos los 50 estados; después se eligen al azar ciudades y pueblos en esos condados; luego aleatoriamente se seleccionan cuadras residenciales en cada ciudad o pueblo; luego se escogen hogares al azar en cada cuadra y, por último, se selecciona al azar a una persona de cada hogar. En este libro no utilizaremos un diseño muestral de este tipo. Hay que recalcar otra vez que los métodos de este libro por lo regular requieren una *muestra aleatoria simple*.

Errores de muestreo

Por muy bien que usted planee y ejecute el proceso de recolección de muestras, es probable que ocurra algún error en los resultados. Por ejemplo, seleccione a 1000 adultos al azar, pregúnteles si se graduaron de bachillerato y registre el porcentaje de respuestas afirmativas en la muestra. Si selecciona otra muestra de 1000 adultos al azar, es probable que obtenga un porcentaje *diferente* en esa muestra.

Definiciones

Un **error de muestreo** es la diferencia entre el resultado de una muestra y el verdadero resultado de la población; tal error es consecuencia de las posibles fluctuaciones de las muestras.

Un **error no de muestreo** ocurre cuando los datos de una muestra se obtienen, registran o analizan de forma incorrecta (como cuando se selecciona una muestra sesgada o predispuesta, cuando se usa un instrumento de medición defectuoso o cuando se cometen errores al copiar los datos).

Si recolectamos con cuidado una muestra que sea representativa de la población, usaremos los métodos de este libro para analizar el error de muestreo, pero debemos tener sumo cuidado para minimizar el error no de muestreo.

Después de leer esta sección, es normal estar un poco abrumado por la variedad de las diferentes definiciones. Sin embargo, recuerde este punto principal: el método usado para reunir datos es sumamente importante y debemos reconocer que la *aleatoriedad* es importante en particular. Si los datos muestrales no se reúnen de manera adecuada, resultarán inútiles por completo, de forma que ninguna cantidad de tortura estadística pueda salvarlos.

1-4 Destrezas y conceptos básicos

En los ejercicios 1 a 4, determine si la distribución dada corresponde a un estudio observacional o a un experimento.

1. Prueba de fármacos A los pacientes se les administra Lipitor para determinar si este fármaco tiene el efecto de disminuir los altos niveles de colesterol.

2. Tratamiento para la sífilis Ha surgido una gran controversia en torno del estudio de pacientes con sífilis que no recibieron un tratamiento que los habría curado. Su salud fue vigilada por años después de que se descubrió que padecían sífilis.

3. Fraude al consumidor El departamento de pesos y medidas del condado de Dutchess selecciona al azar expendios de gasolina y obtiene un galón de gasolina de cada bomba. La cantidad bombeada se mide para comprobar la precisión.

4. Brazaletes magnéticos A los pasajeros de un barco de crucero se les dan brazaletes magnéticos, que aceptan usar en un intento por disminuir o eliminar los efectos del mareo.

En los ejercicios 5 a 8, identifique el tipo de estudio observacional (transeccional, retrospectivo o prospectivo).

5. Investigación médica Un investigador de la escuela de medicina de la Universidad de Nueva York obtiene datos acerca de heridas en la cabeza examinando los registros del hospital de los últimos cinco años.

6. Psicología del trauma Un investigador del hospital Monte Sinaí, en la ciudad de Nueva York, planea obtener datos haciendo seguimiento (hasta el año 2010) a los hermanos de las víctimas que perecieron en el ataque terrorista al World Trade Center el 11 de septiembre de 2001.

7. Estadísticas de desempleo El Departamento de Trabajo de Estados Unidos obtiene datos de desempleo reales encuestando a 50,000 personas en este mes.

8. Ganadores de la lotería Un economista reúne datos entrevistando a personas que ganaron la lotería entre los años 1995 y 2000.

En los ejercicios 9 a 20, identifique cuál de estos tipos de muestreo se utiliza: aleatorio, sistemático, por conveniencia, estratificado o por racimos.

9. Noticias televisivas Un reportero de noticias de la empresa de televisión NBC pretende conocer la reacción a una historia triste entrevistando a las personas que van pasando frente a su estudio.

10. Selección de jurado El comisionado de jurados del condado de Dutchess obtiene una lista de 42,763 propietarios de automóviles y compone una junta de jurados seleccionando cada 100-ésimo nombre en esa lista.

11. Encuestas telefónicas En una encuesta de Gallup de 1059 adultos, los sujetos entrevistados fueron seleccionados mediante el uso de una computadora, para generar aleatoriamente los números telefónicos a los que se llamó.

12. Propiedad de automóviles Una investigadora de General Motors dividió todos los automóviles registrados en categorías de subcompacto, compacto, mediano, intermedio y grande. Ella encuesta a 200 propietarios de automóviles de cada categoría.

13. Estudiantes que beben La Universidad de Newport, motivada por un estudiante que murió en estado de ebriedad, realizó una investigación de estudiantes que beben seleccionando al azar 10 diferentes salones de clase y entrevistando a todos los estudiantes en cada uno de estos grupos.

14. Marketing Una ejecutiva de marketing de General Motors encontró que su departamento de relaciones públicas acababa de imprimir sobres con los nombres y direcciones de todos los propietarios de un Corvette. Ella quiere hacer una prueba piloto de la nueva estrategia de mercadotecnia, así que mezcla cuidadosamente todos los sobres en una urna y obtiene un grupo de muestra sacando 50 de esos sobres.

15. Puesto de revisión de sobriedad El autor fue un observador en un puesto de revisión de sobriedad de la policía donde se detenía y entrevistaba a cada quinto conductor. (El autor fue testigo del arresto de un ex alumno).

16. Encuestas de salida La CNN está planeando una encuesta de salida en que se elegirán aleatoriamente 100 casillas electorales y todos los votantes se entrevistarán conforme vayan saliendo de los locales.

17. La educación y el salario Un economista estudia el efecto de la educación en el salario, y realiza una encuesta a 150 trabajadores seleccionados al azar de cada una de estas categorías: estudios menores que la secundaria, grado de escuela secundaria, estudios de mayor grado que la secundaria.

18. Antropometría Un estudiante de estadística obtuvo datos de estatura/peso entrevistando a los miembros de la familia.

19. Investigación médica Un investigador de la Universidad Johns Hopkins encuesta a todos los pacientes cardiacos en cada uno de 30 hospitales seleccionados al azar.

20. Encuesta de MTV Un experto en marketing está planeando una encuesta para MTV, en la cual 500 personas se elegirán aleatoriamente de cada grupo de edades de 10 a 19, 20 a 29, etcétera.

Los ejercicios 21 a 26 se relacionan con muestras aleatorias y muestras aleatorias simples.

21. Muestreo de tabletas de aspirina Un farmacéutico mezcla cuidadosamente un recipiente con 1000 tabletas de Bufferin y luego recoge una muestra de 50 tabletas que se evaluarán para determinar el contenido exacto de aspirina. ¿Este plan de muestreo describe un muestreo aleatorio? ¿Un muestreo aleatorio simple? Explique.

22. Muestreo de estudiantes Un salón de clases consta de 30 estudiantes sentados en cinco filas diferentes, con seis estudiantes en cada fila. El profesor tira un dado y el resultado se utiliza para seleccionar una muestra de los estudiantes de una fila particular. ¿Este plan de muestreo es un muestreo aleatorio? ¿Un muestreo aleatorio simple? Explique.

23. Muestreo de conveniencia Un reportero de noticias se para en la esquina de una calle, obtiene una muestra de residentes de la ciudad seleccionando a cinco adultos que pasan por ahí y les pregunta acerca de sus hábitos de fumar. ¿Este plan de muestreo dará como resultado un muestreo aleatorio? ¿Un muestreo aleatorio simple? Explique.

24. Muestreo sistemático Un ingeniero de control de calidad selecciona cada 100-ésima unidad de fuente de poder de computadora que pasa por una banda transportadora. ¿Resulta este plan de muestreo en un muestreo aleatorio? ¿Un muestreo aleatorio simple? Explique.

25. Muestra estratificada La empresa de alimentos General Foods planea realizar una encuesta de marketing a 100 hombres y 100 mujeres en el condado de Orange, que consiste en un número igual de hombres y mujeres. ¿Resulta este plan de muestreo en un muestreo aleatorio? ¿Un muestreo aleatorio simple? Explique.

26. Muestra por racimos Un investigador de marketing selecciona aleatoriamente 10 cuadras en el pueblo de Newport, luego pregunta a los adultos residentes de las cuadras seleccionadas si tienen un reproductor de DVD. ¿Este plan de muestreo resultará en un muestreo aleatorio? ¿Un muestreo aleatorio simple? Explique.

1-4 Más allá de lo básico

27. Diseño de muestreo La compañía de publicaciones Addison-Wesley le ha comisionado a usted para encuestar a 100 estudiantes usuarios de esta obra. Describa los procedimientos para obtener una muestra de cada tipo: aleatoria, sistemática, de conveniencia, estratificada y por racimos.

28. Confusión Mencione un ejemplo (diferente del que está en el texto) que ilustre la forma en que ocurre la confusión.

29. Selección aleatoria Entre las 50 entidades de Estados Unidos, se elige aleatoriamente una entidad. Después se obtiene el padrón electoral de todo el estado y se selecciona un nombre al azar. ¿Este procedimiento resultará en un votante seleccionado aleatoriamente?

30. Diseño muestral En el artículo "Cardiovascular Effects of Intravenous Triiodothyronine in Patients Undergoing Coronary Artery Bypass Graft Surgery" (*Journal of the American Medical Association*, vol. 275, núm. 9), los autores explican que los pacientes fueron asignados a uno de tres grupos: **1.** un grupo tratado con triyodotironina, **2.** un grupo tratado con una píldora de sal normal y dopamina, y **3.** un grupo placebo al que se le dio una píldora de sal común. Los autores resumen el diseño muestral como un "experimento prospectivo, aleatorizado, doble ciego, placebo y controlado". Describa el significado de cada uno de estos términos en el contexto de este estudio.

31. Conductores con teléfonos celulares ¿Cuáles son los dos problemas principales que pueden encontrarse en un estudio prospectivo, donde algunos conductores no tienen teléfonos celulares mientras que a otros se les pide que usen sus teléfonos celulares mientras conducen?

Repaso

Este capítulo presentó algunos fundamentos importantes, consistentes en definiciones básicas, como las de *muestra* y *población*, junto con algunos principios esenciales. La sección 1-2 analizó los diferentes tipos de datos. La sección 1-3 trató con el uso del pensamiento crítico en el análisis y la evaluación de resultados estadísticos. La sección 1-4 introdujo elementos importantes en el diseño de experimentos. Al terminar el estudio de este capítulo, usted debe ser capaz de:

- Distinguir entre una población y una muestra, y entre un parámetro y un estadístico.
- Identificar el nivel de medición (nominal, ordinal, de intervalo, de razón) de un conjunto de datos.
- Entender la importancia de un buen diseño experimental, incluyendo el control de los efectos de las variables, la réplica y la aleatorización.
- Reconocer la importancia de seguir buenos métodos de muestreo en general y reconocer la importancia de una *muestra aleatoria simple* en particular. Entender que si los datos muestrales no se reúnen de manera adecuada, los datos resultarían tan inútiles que ninguna cantidad de tortura estadística podría salvarlos.

Ejercicios de repaso

1. Muestreo Poco después de que las torres del World Trade Center se colapsaran por los ataques terroristas, America Online aplicó una encuesta a sus suscriptores de Internet y preguntó lo siguiente: "¿Deben reconstruirse las torres del World Trade Center?". De 1,304,240 personas que respondieron, 768,731 respondieron "sí", 286,756 contestaron "no", y 248,753 dijeron que era "demasiado pronto para decidir". Como esta muestra es extremadamente grande, ¿se puede considerar que las respuestas sean representativas de la población de Estados Unidos? Explique.

2. Diseño de muestreo Usted ha sido contratado por Visa para realizar un estudio acerca del uso de tarjeta de crédito entre los estudiantes de tiempo completo que asisten a su universidad. Describa un procedimiento para obtener una muestra de cada tipo: aleatoria, sistemática, de conveniencia, estratificada y por racimos.

3. Identifique el nivel de medición (nominal, ordinal, de intervalo, de razón) que se aplica a cada uno de los siguientes ejemplos.

a. Los pesos de las personas en una muestra de pasajeros de un elevador.

b. Una clasificación de crítica de cine de *debe verse*, *recomendada, no recomendada*, *ni piense en verla*.

c. Bob, que es distinto en muchas formas, mide el tiempo en días a partir de 0, que corresponde a su fecha de nacimiento. El día anterior a su nacimiento es -1, el día después de su nacimiento es +1, etcétera. Bob ha convertido fechas de eventos históricos importantes a su sistema de numeración. ¿Cuál es el nivel de medición de estos números?

4. Coca Cola La Coca Cola Company tiene 366,000 accionistas y efectúa una encuesta por medio de la selección aleatoria de 30 accionistas de cada uno de los 50 estados de Estados Unidos. Se registra el número de acciones de cada accionista de la muestra.

a. ¿Los valores obtenidos son discretos o continuos?
b. Identifique el nivel de medición (nominal, ordinal, de intervalo, de razón) de los datos muestrales.
c. ¿Qué tipo de muestreo (aleatorio, sistemático, de conveniencia, estratificado, por racimos) se utiliza?
d. Si se calcula el número promedio (la media) de acciones, ¿el resultado es un estadístico o un parámetro?
e. Si usted fuera el ejecutivo en jefe de la Coca Cola Company, ¿qué característica del conjunto de datos consideraría que es extremadamente importante?
f. ¿Qué es lo que está incorrecto al evaluar la opinión del accionista enviando un cuestionario por correo que los accionistas podrían llenar y regresar por el mismo medio?

5. Más Coca Cola Identifique el tipo de muestreo (aleatorio, sistemático, de conveniencia, estratificado, por racimos) que se utiliza cuando una muestra de 366,000 accionistas de Coca Cola se obtiene como ya se describió. Después determine si el esquema de muestreo parece resultar en una muestra representativa de la población de los 366,000 accionistas.

a. Se compila una lista completa de todos los accionistas y se selecciona cada 500-ésimo nombre.
b. En la junta anual de accionistas, se realiza una encuesta de todos los asistentes.
c. Se seleccionan al azar 50 diferentes corredores de bolsa y se hace una encuesta a todos sus clientes que tengan acciones de Coca Cola.
d. Se compila un archivo de computadora de todos los accionistas, de manera que todos ellos se numeran de forma consecutiva y después los números aleatorios generados por computadora se utilizan para seleccionar la muestra de accionistas.
e. Se reúnen todos los códigos postales de los accionistas y se elige al azar a cinco accionistas de cada código postal.

6. Diseño de experimento Usted planea realizar un experimento para probar la eficacia del Sleepeze, un nuevo fármaco que se supone que reducirá el efecto del insomnio. Usará una muestra de sujetos que han sido tratados con el fármaco y otra muestra de sujetos a quienes se les administró un placebo.

a. ¿Qué es el “estudio ciego” y como puede usarse en este experimento?
b. ¿Por qué es importante el uso del estudio ciego en este experimento?
c. ¿Qué es un diseño de bloques completamente aleatorizado?
d. ¿Qué es un diseño de bloques rigurosamente controlado?
e. ¿Que es la réplica y por qué es importante?

Ejercicios de repaso acumulativos

Los ejercicios de repaso acumulativos de este libro, están diseñados para incluir temas de capítulos anteriores. Para los capítulos 2 a 13, estos ejercicios incluyen temas de capítulos anteriores. Para este capítulo presentamos ejercicios de calentamiento para calculadora con expresiones similares a las que se encuentran a lo largo de esta obra. Utilice su calculadora para obtener los valores indicados.

1. Remítase al conjunto de datos 1 del apéndice B y considere sólo los pesos de los primeros 10 varones. ¿Qué valor se obtiene cuando se suman estos 10 pesos y el total se divide entre 10? (Este resultado, llamado media, se analiza en el capítulo 2).

2. $\dfrac{98.20 - 98.60}{0.62}$

3. $\dfrac{98.20 - 98.60}{\dfrac{0.62}{\sqrt{106}}}$

4. $\left[\dfrac{1.96 \cdot 15}{2}\right]^2$

5. $\sqrt{\dfrac{(5-7)^2 + (12-7)^2 + (4-7)^2}{3-1}}$

6. $\dfrac{(183 - 137.09)^2}{137.09} + \dfrac{(30 - 41.68)^2}{41.68}$

7. $\sqrt{\dfrac{10(513.27) - 71.5^2}{10(10-1)}}$

8. $\dfrac{8(151{,}879) - (516.5)(2176)}{\sqrt{8(34{,}525.75) - 516.5^2}\ \sqrt{8(728{,}520) - 2176^2}}$

En los ejercicios 9 a 12, las expresiones dadas están diseñadas para dar resultados expresados en notación científica. Por ejemplo, el resultado de la pantalla de la calculadora de 1.23E5 (o 1.23^5 en algunas calculadoras) puede expresarse como 123,000, y el resultado de 4.65E−4 (o 4.65^{-4} en algunas calculadoras) puede expresarse como 0.000456. Realice la operación que se indica y exprese el resultado como un número ordinario, no en notación científica.

9. 0.95^{500} **10.** 8^{14} **11.** 9^{12} **12.** 0.25^{17}

Actividades de cooperativas en equipo

1. Actividad en clase Obtenga 18 popotes de la cafetería. Corte 6 de ellos por la mitad, corte 6 en cuartos y los otros 6 déjelos como están. Ahora debe haber 42 popotes de diferentes longitudes. Póngalos en una bolsa, revuélvalos, luego seleccione un popote, note su longitud y póngalo de nuevo en la bolsa. Repita esto hasta seleccionar 20 popotes. *Importante:* Seleccione los popotes sin mirar al interior de la bolsa y saque el primero que toque. Calcule el promedio (media) de la muestra de 20 popotes. Ahora saque todos los popotes y encuentre la media de la población. ¿La muestra dio un promedio cercano al promedio de la población real? ¿Por qué sí o por qué no?

2. Actividad en clase A mediados de diciembre de un año reciente, el proveedor de servicios de Internet America Online (AOL) efectuó una encuesta a sus usuarios. Se les preguntó lo siguiente acerca de los árboles de Navidad: “¿Cuál prefiere usted?”. La respuesta podía ser “un árbol natural” o “un árbol artificial”. Entre las 7073 respuestas recibidas de los usuarios de Internet, 4650 preferían un árbol natural y 2423 un árbol artificial. Ya señalamos que como la muestra es una muestra de respuesta voluntaria, no es posible obtener conclusiones acerca de una población mayor que las 7073 personas que respondieron. Identifique otros problemas en esta pregunta de encuesta.

3. Actividad en clase Identifique los problemas en los siguientes eventos:

   - Un reporte televisado recientemente por *CNN Headline News* incluyó el comentario de que el crimen en Estados Unidos disminuyó en la década de 1980 debido al incremento de abortos en la década de 1970, que resultó en un menor número de niños no deseados.
   - La revista *Consumer Reports* envió por correo un cuestionario anual acerca de automóviles y otros productos de consumo. También se incluyó la petición de una contribución económica voluntaria y una votación para el consejo de administración de la revista. Las respuestas debían enviarse por correo en sobres que requerían timbres postales.

Proyecto tecnológico

El propósito de este proyecto es introducir los recursos tecnológicos que usted usará en su curso de estadística. Remítase al conjunto de datos 14 en el apéndice B y use las puntuaciones de facilidad de lectura de Flesch para el libro *Harry Potter y la piedra filosofal* de J. K. Rowling. Utilizando su programa de estadística o la calculadora TI-83 Plus, introduzca estos 12 valores, y luego imprima un listado de ellos.

STATDISK Haga clic en **Data (datos)** en la parte superior de la pantalla, después seleccione **Sample Editor (editor de muestra)** y proceda a introducir los datos. Para imprimir seleccione con el ratón **File (archivo)** y luego seleccione **Print (imprimir)**.

Minitab Introduzca los datos en la columna C1, después haga clic en **File (archivo)** y seleccione **Print Worksheet (imprimir hoja de cálculo)**.

Excel Introduzca los datos en la columna A, después haga clic en **File (archivo)** y seleccione **Print (imprimir)**.

TI-83 Plus La impresión de la pantalla de la TI-83 Plus sólo es posible mediante el uso de la conexión a una computadora *Graphlink*.

de los DATOS a la DECISIÓN

Pensamiento crítico

El médico suizo H. C. Lombard en una ocasión compiló datos sobre la longevidad en relación con distintas profesiones. Usó actas de defunción que incluían nombre, edad al morir y profesión. Después procedió a calcular el promedio (media) de la longevidad para las diferentes profesiones, y encontró que los estudiantes eran los más bajos con una media de ¡sólo 20.7 años! (Véase "A Selection of Selection Anomalies" de Wainer, Palmer y Bradlow en *Chance*, vol. 11, núm. 2). Si los mismos datos fueran reunidos el día de hoy en Estados Unidos, se obtendrían resultados similares.

Análisis de los resultados

¿En realidad ser estudiante es más peligroso que ser agente de policía, chofer de taxi o empleado postal? Explique.

PROYECTO DE INTERNET

El sitio Web de *Estadística*/Triola

En esta sección de cada capítulo, se le pedirá que visite la página Web de este libro. Desde ahí usted puede tener acceso a las páginas referentes a todos los proyectos de Internet que vienen en la *novena edición* de *Estadística*/Triola. Vaya a este sitio ahora y familiarícese con todas las características de este libro a las que tiene acceso.

Cada proyecto de Internet incluye actividades, como la exploración de conjuntos de datos, la ejecución de modelos de simulación y la investigación de ejemplos de la vida real, que se encuentran en varios sitios Web. Estas actividades le ayudarán a explorar y entender la rica naturaleza de la estadística y su importancia en nuestro mundo. ¡Visite el sitio del libro ahora y disfrute de las exploraciones!

www.pearsoneducacion.net/triola

La estadística @ en el trabajo

"Empleamos la estadística para determinar el grado de aislamiento que existe entre grupos putativos".

Sarah Mesnick

Ecologista conductual y molecular

Sara Mesnick es miembro posdoctorado del National Research Council. Su trabajo como bióloga en mamíferos incluye no sólo investigación en el mar, sino también en el Laboratory of Molecular Ecology. Sus estudios se enfocan en la organización social y estructura poblacional de los cachalotes. Obtuvo su doctorado en biología evolutiva en la Universidad de Arizona.

¿A qué se dedica?

Mi investigación se enfoca en la relación que existe entre la sociabilidad y la estructura poblacional de los cachalotes. Nosotros empleamos esta información para crear mejores modelos de manejo para la conservación de ésta y otras especies de mamíferos marinos en peligro de extinción.

¿Qué conceptos de la estadística utiliza?

En la actualidad utilizo la chi cuadrada y el estadístico F para examinar la estructura poblacional, y medidas de regresión para estimar el grado de relación entre los individuos de la manada de ballenas. Empleamos la chi cuadrada y el estadístico F para determinar la cantidad de poblaciones discretas de ballenas en el Pacífico. Estas poblaciones se manejan como grupos independientes. El análisis de regresión de la relación se utiliza para determinar el parentesco dentro de los grupos.

¿Podría citar un ejemplo específico que ilustre el uso de la estadística?

Actualmente estoy trabajando con muestras de tejido que obtengo de tres encallamientos masivos de cachalotes. Utilizamos marcadores genéticos para determinar el grado de parentesco entre los individuos encallados. Se trata de un comportamiento sorprendente: manadas completas nadaron hacia la playa siguiendo a un ballenato hembra, encallaron y después murieron. Pensamos que para hacer algo tan extremo como esto, los individuos implicados debieron tener una relación muy cercana; sin embargo, estamos descubriendo que no es así. La estadística nos permite determinar la probabilidad de que dos individuos sean parientes, dado el número de alelos que comparten. Además, el cachalote y muchas otras especies de mamíferos marinos, aves y tortugas se lastiman o mueren incidentalmente en maniobras de pesca. Necesitamos conocer el tamaño de la población de la que provienen estos animales; si la población es pequeña y las muertes incidentales abundantes, la población de mamíferos marinos estaría amenazada. Empleamos la estadística para determinar el grado de aislamiento que existe entre grupos putativos. Si resultara que los grupos están aislados, usaríamos esta información para preparar planes de manejo diseñados específicamente para conservar a los mamíferos marinos de la región. Tal vez sean necesarias actividades humanas que protejan la salud del ambiente marino y a sus habitantes.

¿De qué forma enfoca su investigación?

Tratamos de evitar ideas preconcebidas acerca de la forma en que los animales están distribuidos en su medio ambiente. Puesto que los mamíferos marinos en particular son tan difíciles de estudiar, suelen existir ideas aceptadas sobre lo que estos animales hacen, aun cuando esto no se ha investigado de manera profunda. En lo que se refiere al parentesco entre individuos dentro de grupos de cachalotes, alguna vez se pensó que éste era matrilineal y que incluía a un "líder del harem". Con el advenimiento de la tecnología genética, la dedicación en el trabajo de campo, mentes más abiertas y análisis más críticos (aquí interviene la estadística), somos capaces de examinar de nuevo estas ideas.

2

Descripción, exploración y comparación de datos

PROBLEMA DEL CAPÍTULO

¿Realmente son afectadas las personas que no fuman cuando otros sí lo hacen junto a ellas? ¿O es un mito el efecto del *fumador pasivo*?

El conjunto número 6 del Apéndice B incluye algunos de los datos disponibles más recientes del National Institute of Health de Estados Unidos. Los datos, que se reproducen en la tabla 2-1, se obtuvieron como parte del National Health and Nutrition Examination Survey. Los valores de los datos corresponden a los niveles medidos de cotinina sérica (en ng/ml) en personas seleccionadas como sujetos de estudio (los datos se redondearon hacia el entero más cercano, de tal modo que un valor de cero no necesariamente implica la ausencia total de cotinina. De hecho, todos los valores originales fueron mayores que cero). La cotinina es un metabolito de la nicotina, es decir, es una sustancia que se produce cuando el cuerpo absorbe la nicotina. Porque se sabe que la nicotina se absorbe cuando se consumen cigarrillos, hay una forma indirecta de medir la presencia efectiva del humo del tabaco; esto es, por medio de la cotinina.

Existen varios aspectos importantes al respecto: ¿deben preocuparse por su salud las personas que no fuman ante la presencia de fumadores activos? Para preverlo, en los últimos años las autoridades sanitarias han elaborado muchos reglamentos para restringir el tabaquismo en lugares públicos. ¿Son justificadas dichas regulaciones por razones de salud o sólo provocan dificultades innecesarias a los fumadores?

Pensamiento crítico: Una comparación visual de las cifras en los tres grupos de la tabla 2-1 proporciona cierta información. En este capítulo presentamos métodos para lograr una mayor comprensión. Seremos capaces de producir comparaciones productivas e inteligentes; aprenderemos técnicas para describir, explorar y comparar conjuntos de datos, tales como los tres grupos de la tabla 2-1.

Tabla 2-1 Niveles medidos de cotinina en tres grupos

Fumador: Los sujetos reportan su consumo de tabaco.

HTA: Humo de tabaco ambiental). Sujetos que no fuman, pero que están expuestos a humo de tabaco ambiental (“fumadores pasivos”), en su casa o trabajo.

SHTA: (Sin humo de tabaco ambiental): Sujetos que no fuman y que no se exponen a humo de tabaco ambiental en su casa o trabajo. Esto es, no fuman ni son fumadores pasivos.

Fumador:	1	0	131	173	265	210	44	277	32	3
	35	112	477	289	227	103	222	149	313	491
	130	234	164	198	17	253	87	121	266	290
	123	167	250	245	48	86	284	1	208	173
HTA:	384	0	69	19	1	0	178	2	13	1
	4	0	543	17	1	0	51	0	197	3
	0	3	1	45	13	3	1	1	1	0
	0	551	2	1	1	1	0	74	1	241
SHTA:	0	0	0	0	0	0	0	0	0	0
	0	9	0	0	0	0	0	0	244	0
	1	0	0	0	90	1	0	309	0	0
	0	0	0	0	0	0	0	0	0	0

2-1 Panorama general

Este capítulo es sumamente importante, ya que presenta las herramientas básicas para medir y describir diferentes características de un conjunto de datos. Cuando se describen, exploran y comparan conjuntos de datos, las siguientes características suelen ser de enorme importancia.

Características importantes de los datos

1. **Centro:** Valor representativo o promedio que indica la localización de la mitad del conjunto de los datos.
2. **Variación:** Medida de la cantidad en que los valores de los datos varían entre sí.
3. **Distribución:** Naturaleza o forma de la distribución de los datos (tales como normales, uniformes o sesgadas).
4. **Datos distantes:** Valores muestrales que están muy alejados de la vasta mayoría de los demás valores de la muestra.
5. **Tiempo:** Características cambiantes de los datos a través del tiempo.

Sugerencia: La memorización suele ser ineficaz para recordar información importante. Sin embargo, las cinco características anteriores son tan importantes que deben recordarse con el uso de una técnica mnemónica conocida por las iniciales CVDDT; por ejemplo, "**C**uidado con los **V**irus que **D**estruyen **D**atos y **T**rabajo". Se ha visto que dichas técnicas de memorización son muy efectivas para recordar importantes palabras clave que evocan conceptos básicos.

Pensamiento crítico e interpretación: más allá de las fórmulas

Los profesores de estadística, por lo general, piensan que no es tan importante memorizar fórmulas o realizar cálculos aritméticos complejos a mano. Por el contrario, suelen enfocarse en la obtención de resultados por medio del uso de algún tipo de herramienta tecnológica (calculadoras o programas de cómputo), para después entender, de forma práctica, los resultados a través del pensamiento crítico. Tenga esto en mente conforme avance en el estudio de este capítulo. Por ejemplo, cuando estudie la muy importante *desviación estándar*, en la sección 2-5, trate de observar por qué la fórmula clave funciona como una medida de variación, después aprenda a calcular los valores de las desviaciones estándar, pero trabaje realmente en la *comprensión* y la *interpretación* de los valores de la desviación estándar.

Aun cuando este capítulo incluye, de forma detallada, los casos para procedimientos importantes, no es necesario conocerlos a la perfección en todas las situaciones. No obstante, recomendamos que, en cada caso, realice algunos cálculos manuales antes de utilizar su calculadora o computadora. Lo anterior hará que su comprensión se incremente y podrá apreciar mejor los resultados obtenidos con las herramientas tecnológicas.

Los métodos de este capítulo suelen denominarse métodos de **estadística descriptiva**, porque su objetivo es resumir o *describir* las características importantes de un conjunto de datos. Más adelante, utilizaremos métodos de **estadística inferencial**; lo haremos cuando usemos datos muestrales para hacer inferencias (o generalizaciones) acerca de una población. Con la estadística inferencial realizamos una deducción que va más allá de los datos conocidos. La materia de estadística tiene dos divisiones generales: la descriptiva y la inferencial; este capítulo trata los conceptos básicos de la estadística descriptiva.

2-2 Distribuciones de frecuencias

Cuando se trabaja con conjuntos grandes de datos, con frecuencia es útil organizarlos y resumirlos por medio de la construcción de una tabla que liste los distintos valores posibles de los datos (ya sea de forma individual o por grupos), junto con las frecuencias correspondientes, es decir, el número de veces que ocurren dichos valores.

Definición

Distribución de frecuencias: lista valores de datos (ya sea de manera individual o por grupos de intervalos), junto con sus frecuencias (o conteos) correspondientes.

La tabla 2-2 es una distribución de frecuencias que resume los niveles medidos de cotinina de los 40 fumadores que se muetran en la tabla 2-1. La **frecuencia** de una clase particular es el número de valores originales que caen dentro de esa clase. Por ejemplo, la primera clase de la tabla 2-2 tiene una frecuencia de 11, lo que indica que 11 de los valores originales de los datos están entre 0 y 99, inclusive.

Para empezar, presentaremos algunos términos estándar utilizados al referirse a la distribución de frecuencia; después describiremos la forma en que se construyen e interpretan.

Tabla 2-2 Distribución de frecuencias de los niveles de cotinina de los fumadores

Cotinina	Frecuencia
0–99	11
100–199	12
200–299	14
300–399	1
400–499	2

Definiciones

Los **límites de clase inferiores** son las cifras más pequeñas que pueden pertenecer a las diferentes clases. (Los límites de clase inferiores de la tabla 2-2 son 0, 100, 200, 300 y 400).

Los **límites de clase superiores** son las cifras más grandes que pueden pertenecer a las diferentes clases. (Los límites de clase superiores de la tabla 2-2 son 99, 199, 299, 399 y 499).

Las **fronteras de clase** son las cifras utilizadas para separar las clases, aunque sin los espacios creados por los límites de clase. Se obtienen de la siguiente manera: se determina el tamaño del espacio entre el límite de clase superior de una clase y el límite de clase inferior de la siguiente. Se suma la mitad de esa cantidad a cada límite de clase superior, para obtener las fronteras de clase superiores; se resta la mitad de esa cantidad de cada límite de clase inferior, para obtener las fronteras de clase inferiores. (Los espacios de la tabla 2-2 son exactamente de una unidad, de modo que a los límites de clase superiores se les suma 0.5 y a los límites de clase inferiores se les resta 0.5. Las fronteras de la primera clase son −0.5 y 99.5, en tanto que las de la segunda clase son 99.5 y 199.5, y así

continúa

Gráficas de crecimiento actualizadas

Los pediatras acostumbran utilizar gráficas de crecimiento estandarizadas para comparar el peso y la altura de sus pacientes con una muestra de otros niños. Se considera que los pequeños están dentro de un rango normal si su peso y su estatura caen dentro de los percentiles 5o y 95o. Si se encuentran fuera de este rango, les aplicarán pruebas para estar seguros de que no hay problemas médicos graves. Los pediatras se dan cuenta de la existencia de un problema importante gracias a las gráficas; debido a que se basan en niños que vivieron entre 1929 y 1975, se concluyó que las gráficas de crecimiento eran imprecisas. Para rectificar tal problema, las gráficas se actualizaron en el año 2000, con la finalidad de que reflejaran las mediciones actuales de millones de niños. Los pesos y las estaturas de los niños son buenos ejemplos de poblaciones que cambian con el paso del tiempo. Ésta es la razón que lleva a considerar las características cambiantes de los datos a lo largo del tiempo como un aspecto importante de una población.

sucesivamente. La lista completa de fronteras utilizadas para todas las clases es la siguiente: −0.5, 99.5, 199.5, 299.5, 399.5 y 499.5.)

Las **marcas de clase** son los puntos medios de las clases. (Las marcas de clase de la tabla 2-2 son 49.5, 149.5, 249.5, 349.5 y 449.5). Cada marca de clase se calcula sumando el límite de clase inferior con el límite de clase superior y dividiendo la suma entre dos.

La **anchura de clase** es la diferencia entre dos límites de clase inferiores consecutivos o dos fronteras de clase inferiores consecutivas. (La anchura de clase que se utiliza en la tabla 2-2 es igual a 100).

Las definiciones de anchura de clase y frontera de clase son engañosas. Hay que tener cuidado para evitar el error común de considerar la amplitud de clase como la diferencia entre el límite de clase inferior y el límite de clase superior. Vea la tabla 2.2 y observe que la anchura de clase es de 100 y no de 99. El proceso para determinar las fronteras de clase se simplifica si se comprende que éstas básicamente llenan los espacios entre clases al dividir la diferencia entre el final de una clase y el inicio de la siguiente.

Procedimiento de construcción de una distribución de frecuencias

Las distribuciones de frecuencias se construyen por las siguientes razones: **1.** es posible resumir conjuntos grandes de datos, **2.** se logra cierta comprensión respecto de la naturaleza de los datos, y **3.** se llega a tener un avance para construir gráficas importantes (tales como *histogramas*, que se presentarán en la siguiente sección). Muchas de las herramientas tecnológicas permiten obtener de manera automática las distribuciones de frecuencias, sin necesidad de tenerlas que construir manualmente; no obstante, a continuación se presenta el procedimiento básico:

1. Decida el número de clases que desea tener. Debe ser de entre 5 y 20, y deben utilizarse números enteros o redondeados.
2. Calcule

$$\text{Anchura de clase} \approx \frac{(\text{valor más alto}) - (\text{valor más bajo})}{\text{número de clases}}$$

   Redondee el resultado para obtener un número más adecuado (generalmente se redondea hacia *arriba*). Es probable que necesite cambiar el número de clases, pero la prioridad debe ser utilizar valores que sean fáciles de comprender.
3. Punto de partida: comience por elegir un número para el límite inferior de la primera clase. Elija el valor del dato más bajo o un valor conveniente que sea un poco más pequeño.
4. Con el uso del límite más bajo de la primera clase y la anchura de clase, proceda a listar los demás límites de clase inferior. (Sume la anchura de clase al punto de partida para obtener el segundo límite de clase inferior. Después, sume la anchura de clase al segundo límite de clase inferior para obtener el tercero y así sucesivamente).
5. Anote los límites inferiores de clase en una columna vertical y luego proceda a anotar los límites superiores de clase, que pueden identificarse con facilidad.

6. Ponga una marca en la clase apropiada para cada dato. Utilice las marcas para obtener la frecuencia total de cada clase.

Cuando construya una distribución de frecuencias, asegúrese de que las clases no se traslapen, de modo que cada uno de los valores originales pertenezca exactamente a una de las clases. Incluya todos los casos, aun aquellos que tienen una frecuencia de cero. Trate de utilizar la misma anchura para todas las clases, aunque en ocasiones es imposible evitar los intervalos con finales abiertos, como "65 años o mayores".

PROBLEMA DEL CAPÍTULO 2

EJEMPLO Niveles de cotinina de fumadores Utilice los 40 niveles de cotinina de los *fumadores* de la tabla 2-1 y siga el procedimiento anterior para crear la distribución de frecuencias que se muestra en la tabla 2-2. Suponga que desea incluir cinco clases.

SOLUCIÓN

Paso 1: Comience seleccionando cinco clases.

Paso 2: Calcule la anchura de clase. En el siguiente cálculo, 98.2 se redondea a 100, ya que es un número más conveniente.

$$\text{anchura de clase} \approx \frac{(\text{valor más alto}) - (\text{valor más bajo})}{\text{número de clases}} = \frac{491 - 0}{5} = 98.2 \approx 100$$

Paso 3: Elija un punto de partida de 0, que es el valor más bajo en la lista y también porque es un número conveniente.

Paso 4: Sume la anchura de clase de 100 al punto de partida de 0 para determinar que el segundo límite inferior de clase es igual a 100. Continúe, y sume la anchura de clase de 100 para obtener los límites inferiores de clase restantes de 200, 300 y 400.

Paso 5: Liste los límites de clase inferiores de forma vertical, como se muestra al margen. Con esta lista se identifican con facilidad los límites de clases superiores correspondientes, tales como 99, 199, 299, 399 y 499.

0–
100–
200–
300–
400–

Paso 6: Una vez identificados los límites inferiores y superiores de cada clase, proceda a trabajar con el conjunto de datos asignando una marca a cada valor. Ya que completó las marcas, súmelas para obtener las frecuencias que se presentan en la tabla 2-2.

Distribución de frecuencias relativas

Una variante importante de la distribución básica de frecuencias utiliza las **frecuencias relativas**, que se obtienen fácilmente dividiendo cada frecuencia de clase entre el total de frecuencias. Una **distribución de frecuencias relativas** incluye los mismos límites de clase que una distribución de frecuencias, pero utiliza las frecuencias relativas en lugar de las frecuencias reales. Las frecuencias relativas, en ocasiones, se expresan como porcentajes.

$$\text{frecuencia relativa} = \frac{\text{frecuencia de clase}}{\text{suma de todas las frecuencias}}$$

En la tabla 2-3, las frecuencias reales de la tabla 2-2 se reemplazaron con las frecuencias relativas correspondientes, expresadas en porcentajes. La primera clase tiene una frecuencia relativa de $11/40 = 0.275$ o de 27.5%, que se redondea a 28%.

Tabla 2-3 Distribución de frecuencias relativas de los niveles de cotinina en fumadores

Cotinina	Frecuencias relativas
0–99	28%
100–199	30%
200–299	35%
300–399	3%
400–499	5%

La segunda clase tiene una frecuencia relativa de $12/40 = 0.3$ o 30%, y así sucesivamente. Si se construye de manera correcta, la suma de las frecuencias relativas debe totalizar 1 (o 100%), con algunas pequeñas discrepancias, que se permiten al redondear los errores. Puesto que 27.5% se redondeó a 28%, y 2.5% se redondeó a 3%, la suma de frecuencias relativas de la tabla 2-3 es de 101%, en lugar de 100%.

Ya que utilizan proporciones simples o porcentajes, las distribuciones de frecuencias nos facilitan la comprensión de la distribución de los datos y nos permiten comparar diferentes conjuntos de datos.

Distribución de frecuencias acumulativas

Otra variante de la distribución de frecuencias estándar se utiliza cuando se buscan totales acumulativos. La **frecuencia acumulativa** de una clase es la suma de las frecuencias para esa clase y todas las clases previas. La tabla 2-4 muestra la distribución de frecuencias acumulativas de la distribución de frecuencias de la tabla 2-2. Con el uso de las frecuencias originales de 11, 12, 14, 1 y 2, sumamos $11 + 12$ para obtener la segunda frecuencia acumulativa de 23; después, sumamos $11 + 12 + 14 = 37$, para obtener la tercera, y así sucesivamente. Vea la tabla 2-4 y observe que, además del uso de frecuencias acumulativas, los límites de clase fueron reemplazados por expresiones como "menor que", las cuales describen el nuevo rango de valores.

Pensamiento crítico: interpretación de las distribuciones de frecuencias

La transformación de datos brutos en una distribución de frecuencias suele ser un medio para un gran fin. Los siguientes ejemplos ilustran la forma en que se utilizan las distribuciones de frecuencias para describir, explorar y comparar conjuntos de datos. (La siguiente sección muestra cómo la elaboración de una distribución de frecuencias suele ser el primer paso en la creación de una gráfica, que presenta la naturaleza de la distribución de forma visual).

EJEMPLO Descripción de los datos Remítase al conjunto de datos 1 en el Apéndice B, que se refieren al pulso de 40 adultos varones que se seleccionaron aleatoriamente. La tabla 2-5 presenta los *últimos dígitos* de tales datos. Si la tasa de pulsaciones se mide contando el número de latidos cardiacos en un minuto, esperamos que los últimos dígitos tengan frecuencias muy similares. Sin embargo, note que la distribución de frecuencias muestra que todos los últimos dígitos son números *pares*; *no* hay números impares. Lo anterior sugiere que las tasas de pulsaciones no se contaron durante un minuto. Tal vez se contaron durante 30 segundos y después se duplicaron los resultados. (Al examinar más las tasas de pulsaciones *originales*, vemos que cada valor original es un múltiplo de cuatro, lo que sugiere que el número de latidos por minuto se contó durante 15 segundos y que después el resultado se multiplicó por cuatro). Es fascinante aprender el método de recolección de datos con la simple descripción de algunas características de los mismos.

Se identifican autores

Entre 1787 y 1788, Alexander Hamilton, John Jay y James Madison publicaron, de forma anónima, el famoso diario *Federalist*, en un intento por convencer a los neoyorquinos de que debían ratificar la Constitución. Se conoció la identidad de la mayoría de los autores de los artículos, pero el autor de 12 de éstos siguió siendo motivo de discusión. Mediante el análisis estadístico del análisis de frecuencias de diversas palabras, ahora concluimos que James Madison es el *probable* autor de esos dos artículos. En muchos de los artículos disputados, la evidencia en favor de la autoría de Madison es abrumadora, al grado de que casi con seguridad afirmamos que estamos en lo correcto.

Tabla 2-4
Distribución de frecuencias acumulativas de los niveles de cotinina en fumadores

Cotinina	Frecuencia Relativa
Menos de 100	11
Menos de 200	23
Menos de 300	37
Menos de 400	38
Menos de 500	40

Tabla 2-5
Últimos dígitos de las tasas de pulsaciones de varones

Último dígito	Frecuencia
0	7
1	0
2	6
3	0
4	11
5	0
6	9
7	0
8	7
9	0

EJEMPLO **Exploración de datos** Para estudiar el comportamiento del géiser Old Faithful, ubicado en el Parque Nacional Yellowstone, los geólogos recolectan datos del tiempo (en minutos) que transcurre entre las erupciones. La tabla 2-6 muestra un resumen de los datos reales obtenidos. Un examen de la distribución de frecuencias reveló un comportamiento inesperado: la distribución del tiempo presenta dos picos distintos. Tal distribución condujo a los geólogos a considerar dos posibles explicaciones.

Tabla 2-6
Tiempo (en minutos) entre las erupciones del géiser Old Faithful

Tiempo	Frecuencia
40–49	8
50–59	44
60–69	23
70–79	6
80–89	107
90–99	11
100–109	1

PROBLEMA DEL CAPÍTULO 2

EJEMPLO **Comparación de conjuntos de datos** El problema que abre este capítulo incluye conjuntos de datos que representan los niveles de cotinina que se midieron en fumadores, en no fumadores expuestos al humo del tabaco y en no fumadores sin exposición al

Tabla 2-7 Niveles de cotinina de los tres grupos

Cotinina	Fumadores	No fumadores expuestos al humo	No fumadores sin exposición al humo
0–99	28%	85%	95%
100–199	30%	5%	0%
200–299	35%	3%	3%
300–399	3%	3%	3%
400–499	5%	0%	0%
500–599	0%	5%	0%

continúa

humo del tabaco. La tabla 2-7 presenta las frecuencias relativas de los tres grupos. Al comparar dichas frecuencias relativas, es claro que la distribución de frecuencias de los fumadores es muy diferente de las de los otros dos grupos. Debido a que los dos grupos de no fumadores (expuestos y no expuestos) tienen una frecuencia tan alta de cantidades de la primera clase, sería útil comparar más esos conjuntos de datos examinando los valores con mayor detalle.

2-2 Destrezas y conceptos básicos

En los ejercicios 1 a 4, identifique la anchura de clase, las marcas de clase y las fronteras de clase para las distribuciones de frecuencias dadas, con base en el conjunto de datos 1 del Apéndice B.

1.

Presión sanguínea sistólica de varones	Frecuencia
90–99	1
100–109	4
110–119	17
120–129	12
130–139	5
140–149	0
150–159	1

2.

Presión sanguínea sistólica de mujeres	Frecuencia
80–99	9
100–119	24
120–139	5
140–159	1
160–179	0
180–199	1

3.

Colesterol en varones	Frecuencia
0–199	13
200–399	11
400–599	5
600–799	8
800–999	2
1000–1199	0
1200–1399	1

4.

Índice de masa corporal de mujeres	Frecuencia
15.0–20.9	10
21.0–26.9	15
27.0–32.9	11
33.0–38.9	2
39.0–44.9	2

En los ejercicios 5 a 8, elabore la distribución de frecuencias relativas que corresponda a la distribución de frecuencias del ejercicio indicado.

5. Ejercicio 1 **6.** Ejercicio 2 **7.** Ejercicio 3 **8.** Ejercicio 4

En los ejercicios 9 a 12, construya la distribución de frecuencias acumulativas que corresponda a la distribución de frecuencias del ejercicio indicado.

9. Ejercicio 1 **10.** Ejercicio 2 **11.** Ejercicio 3 **12.** Ejercicio 4

13. Dado cargado El autor taladró un hoyo en un dado, lo rellenó con plomo y lo lanzó 200 veces. (Sí, el autor tiene mucho tiempo libre). Los resultados se presentan en la distribución de frecuencias al margen. Construya la distribución de frecuencias relativas correspondiente y determine si el dado en cuestión difiere significativamente de un dado que no ha sido “cargado”.

14. Lotería La distribución de frecuencias al margen se basa en los números *Win 4* de la lotería del estado de Nueva York, incluidos en el conjunto de datos 26 del Apéndice B. Elabore la distribución de frecuencias relativas correspondiente y determine si los resultados se seleccionaron de tal forma que todos los dígitos sean igualmente probables.

Tabla del ejercicio 13

Resultado	Frecuencia
1	27
2	31
3	42
4	40
5	28
6	32

Tabla del ejercicio14

Dígito	Frecuencia
0	18
1	12
2	14
3	9
4	17
5	20
6	21
7	26
8	7
9	16

15. Osos Remítase al conjunto de datos 9 del Apéndice B y construya una distribución de frecuencias con los pesos de los osos. Utilice 11 clases, iniciando con el límite de clase inferior de 0, con una anchura de clase de 50 lb.

16. Temperaturas corporales Remítase al conjunto de datos del Apéndice B; después, construya una distribución de frecuencias de las temperaturas corporales para la medianoche del segundo día. Utilice ocho clases, iniciando con el límite de clase inferior de 96.5, con una anchura de clase de 0.4°F. Describa dos características notables del resultado.

17. Circunferencias de cabezas Remítase al conjunto de datos 3 del Apéndice B. Elabore una distribución de frecuencias con las circunferencias de las cabezas de bebés hombres; luego, construya una distribución de frecuencias separada para las circunferencias de las cabezas de los bebés mujeres. En ambos casos, utilice las clases de 34.0–35.9, 36.0–37.9, etcétera. Después compare los resultados y determine si hay una diferencia significativa entre los dos géneros.

18. Películas de dibujos animados para niños Remítase al conjunto de datos 7 del Apéndice B. Construya una distribución de frecuencias con la duración de las escenas de consumo de tabaco que presentan las películas de dibujos animados para niños; luego, elabore una distribución de frecuencias separada con la duración de las escenas en donde se consume alcohol. En ambos casos, utilice las clases de 0–99, 100–199, etcétera. Compare los resultados y determine si hay una diferencia significativa.

19. Corredores del maratón Remítase al conjunto de datos 8 del Apéndice B. Construya una distribución de frecuencias relativas con las edades de la muestra de hombres que terminaron el maratón de la ciudad de Nueva York; después, elabore una distribución de frecuencias relativas separada con las edades de las mujeres. En ambos casos, inicie la primera clase con el límite de clase inferior de 19, con una anchura de clase de 10. Compare los resultados y determine si hay alguna diferencia notable entre los dos grupos.

20. Coca Cola regular/Coca Cola dietética Remítase al conjunto de datos 17 del Apéndice B. Construya una distribución de frecuencias relativas con los pesos de la Coca Cola regular; inicie la primera clase en 0.7900 lb, con una anchura de clase de 0.0050 lb. Después, construya otra distribución de frecuencias relativas con los pesos de la Coca Cola dietética, iniciando la primera clase en 0.7750 lb, con una anchura de clase de 0.0050 lb. Luego, compare los resultados y determine si hay una diferencia significativa. Si es así, dé una posible explicación.

2-2 Más allá de lo básico

21. Interpretación de los efectos de los datos distantes Remítase al conjunto de datos 20 del Apéndice B, de las cargas axiales de latas de aluminio de 0.0111 pulgadas de grosor. A la carga de 504 lb, se le denomina *dato distante*, ya que se encuentra muy lejos del resto de los valores. Construya una distribución de frecuencias que incluya el valor de 504 lb; después, elabore otra sin incluir este valor. En ambos casos, inicie la primera clase en 200 lb y utilice una anchura de clase de 20 lb. Interprete los resultados estableciendo una generalización acerca del efecto que tiene un dato distante en una distribución de frecuencias.

22. Número de clases Los lineamientos de Sturges para la construcción de una distribución de frecuencias sugieren que el número ideal de clases puede aproximarse por medio de $1 + (\log n)/(\log 2)$, donde n es el número de valores de datos. Utilice esta guía para completar la tabla y determine el número ideal de clases.

Tabla del ejercicio 22

Número de valores	Número ideal de clases
16–22	5
23–45	6
	7
	8
	9
	10
	11
	12

2-3 Visualización de los datos

Recuerde que el principal objetivo de este capítulo es aprender técnicas importantes para investigar las características "CVDDT" importantes de los conjuntos de datos: centro, variación, distribución, datos distantes y cambios a lo largo del tiempo. En la sección 2-2 se introdujo la distribución de frecuencias como una herramienta para describir, explorar o comparar *distribuciones* de conjuntos de datos. En esta sección continuaremos el estudio de las distribuciones por medio de la introducción de gráficas, que son dibujos de distribuciones. Conforme avance en esta sección, considere que el objetivo no es simplemente la construcción de gráficas, sino más bien aprender algo acerca de los conjuntos de datos, es decir, comprender la naturaleza de sus distribuciones.

Histogramas

Entre los distintos tipos de gráficas que se presentan en esta sección, el histograma es particularmente importante.

Definición

Histograma es una gráfica de barras en donde la escala horizontal representa clases de valores de datos y la escala vertical representa frecuencias. Las alturas de las barras corresponden a los valores de frecuencia, en tanto que las barras se dibujan de manera adyacente (sin espacios entre ellas).

Es posible construir un histograma tras completar una tabla de distribución de frecuencias para un conjunto de datos. En la figura 2-1 se presentan los niveles de cotinina de fumadores, los cuales corresponden, de forma directa, a la distribución de frecuencias de la tabla 2-2, que se presentó en la sección previa. Cada barra del histograma está marcada con su frontera de clase inferior a la izquierda y su frontera

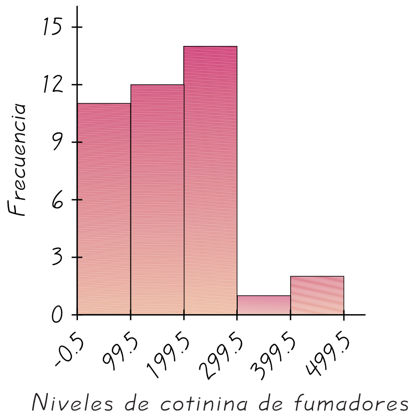


FIGURA 2-1 Histograma

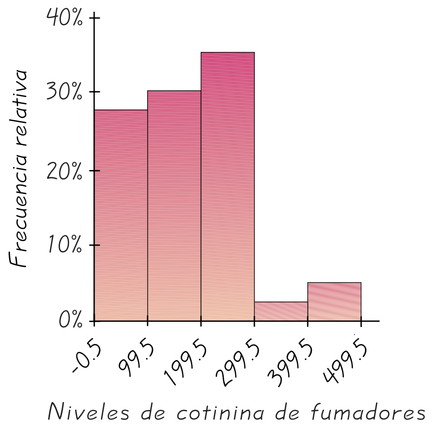


FIGURA 2-2 Histograma de frecuencias relativas

de clase superior a la derecha. En lugar de utilizar fronteras de clase a lo largo de la escala horizontal, suele ser más práctico utilizar los valores de las marcas de clase centradas por debajo de sus barras correspondientes. El uso de los valores de las marcas de clase es muy común en los programas de cómputo que generan histogramas de manera automática.

Antes de construir un histograma, a partir de una distribución de frecuencias completa, debemos mencionar algo acerca de las escalas que se utilizan en los ejes vertical y horizontal. La frecuencia máxima (o el siguiente número conveniente más alto) tiene que sugerir un valor para la parte superior de la escala vertical; el cero habrá de colocarse al inicio. En la figura 2-1 se diseñó una escala vertical que va de 0 a 15. La escala horizontal debe subdividirse de modo tal que permita que se ajusten bien todas las clases. De manera ideal, hay que tratar de seguir la regla práctica del intervalo, la cual establece que la altura vertical del histograma debe medir aproximadamente tres cuartas partes de la anchura total. Ambos ejes tienen que etiquetarse de forma clara.

Interpretación de un histograma Recuerde que el objetivo no es la simple construcción de un histograma, sino aprender algo acerca de los datos. Analice el histograma para ver qué es posible aprender acerca de "CVDDT": el centro de los datos, la variación (que se estudiará en la sección 2.5), la forma de la distribución y la existencia o ausencia de datos distantes (valores que se encuentran lejos de los demás). El histograma no es adecuado para determinar si hay cambios a lo largo del tiempo. Al examinar la figura 2-1, se verá que el histograma se centra alrededor del 175, que los valores varían aproximadamente desde 0 hasta 500 y que la distribución está más cargada hacia la izquierda.

Histograma de frecuencias relativas

Un **histograma de frecuencias relativas** tiene la misma forma y escala horizontal que un histograma, pero la escala vertical está marcada con las frecuencias relativas en lugar de las frecuencias reales, tal como sucede en la figura 2-2.

Polígono de frecuencias

Un **polígono de frecuencias** utiliza segmentos lineales conectados a puntos que se localizan directamente por encima de los valores de las marcas de clase. Véase la figura 2-3 en la página siguiente, que incluye el polígono de frecuencias correspondiente a la tabla 2-2. Las alturas de los puntos corresponden a las frecuencias de clase, en tanto que los segmentos lineales se extienden hacia la derecha y la izquierda, de manera que la gráfica inicia y termina sobre el eje horizontal.

Ojiva

Una ojiva es una gráfica lineal que representa frecuencias *acumulativas*, de la misma forma que la distribución de frecuencias acumulativas es una lista de éstas (véase la tabla 2-4 en la sección anterior). La figura 2-4 es la ojiva correspondiente a la tabla 2-4. Observe que la ojiva utiliza fronteras de clase, a lo largo de la escala horizontal, y que la gráfica empieza con la frontera inferior de la primera clase, en tanto que finaliza con la frontera superior de la última clase. Las ojivas son útiles para determinar el número de valores que se encuentran por debajo de un valor

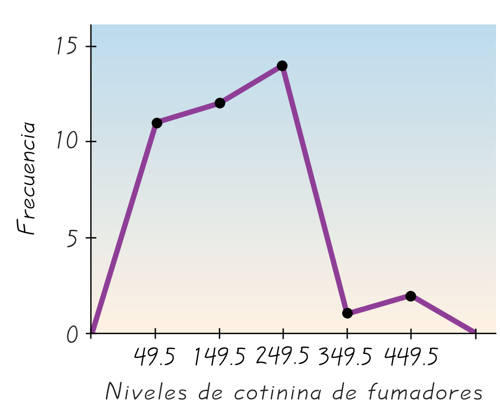


FIGURA 2-3 **Polígono de frecuencias**

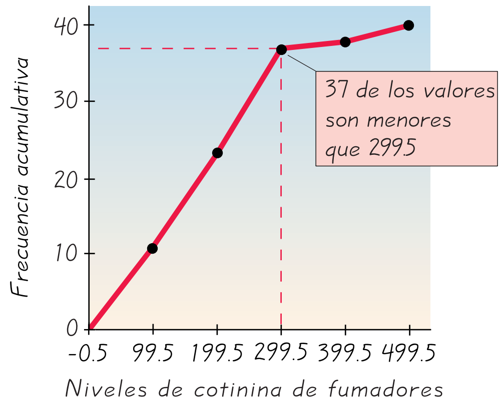


FIGURA 2-4 **Ojiva**

particular. Por ejemplo, la figura 2-4 muestra que 37 de los valores del nivel de cotinina son menores que 299.5.

Gráficas de puntos

Una **gráfica de puntos** consiste en una gráfica en donde se marca cada valor de un dato como un punto a lo largo de una escala de valores. Los puntos que representan valores iguales se amontonan. Observe la figura 2-5, que representa la duración de películas de dibujos animados para niños, que se listan en el conjunto de datos 7 del Apéndice B. Por ejemplo, los dos puntos que aparecen a la izquierda representan el valor de 64 minutos, que ocurre dos veces en el conjunto de datos 7. En esta gráfica de puntos vemos que la duración de 120 minutos difiere mucho de las demás.

Gráficas de tallo y hojas

Una **gráfica de tallo y hojas** representa datos que separan cada valor en dos partes: el tallo (el dígito ubicado en el extremo izquierdo) y la hoja (el dígito del extremo derecho). La ilustración de la siguiente página muestra una gráfica de tallo y hojas de las mismas duraciones de películas listadas en el conjunto de datos 7 del Apéndice B. Dichas duraciones (en minutos), si se acomodan en orden creciente, son 64, 64, 69, 70, 71, 71, 71, 72, 73, . . . , 120. Es fácil ver cómo el primer valor de 64 se separó en su tallo de 6 y su hoja de 4. Cada uno de los valores restantes, lo hace de una manera similar. Note que las hojas se ordenaron en forma creciente y no en el orden en que aparecen en la lista original.

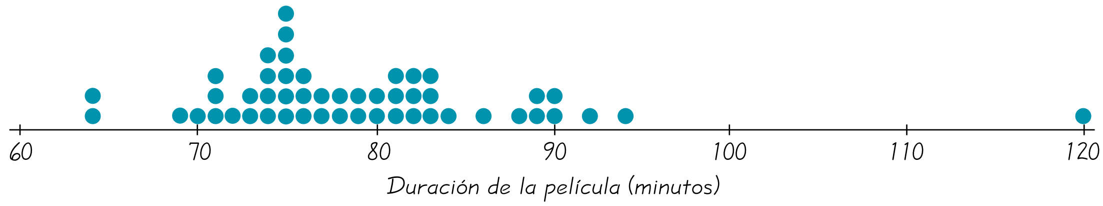


FIGURA 2-5 **Gráfica de puntos de la duración de películas para niños**

Gráfica de tallo y hojas

Tallo (decenas)	Hojas (unidades)	
6	449	←Los valores son 64, 64, 69.
7	01112334444555555666778899	
8	0011122233346899	
9	0024	
10		
11		
12	0	← El valor es 120.

Si colocamos la página de lado, veremos una distribución de tales datos. Una gran ventaja de la gráfica de tallo y hojas radica en que nos permite ver la distribución de los datos y, al mismo tiempo, retener toda la información de la lista original. En caso de ser necesario, reconstruiríamos la lista original de valores. Otra ventaja es que la construcción de una gráfica de tallo y hojas implica una forma fácil y rápida de *ordenar* datos, y algunos procedimientos estadísticos requieren de un ordenamiento (como el cálculo de una mediana o de los percentiles).

Los renglones de datos de una gráfica de tallo y hojas son similares en naturaleza a las barras de un histograma. Uno de los lineamientos para la construcción de histogramas es que se incluyan entre 5 y 20 clases, lo cual se aplica a la gráfica de tallo y hojas por las mismas razones. Por lo general, obtenemos mejores gráficas de tallo y hojas si redondeamos primero los valores de los datos originales. Además, este tipo de gráficas pueden *expandirse* para incluir más renglones y *condensarse* para disminuir el número de renglones. En nuestro ejemplo, la gráfica de tallo y hojas puede expandirse subdividiendo los renglones en otros con hojas que incluyan dígitos del 0 al 4, así como otros con dígitos del 5 al 9, tal como se muestra en el siguiente diagrama.

Gráfica expandida de tallo y hojas

Tallo	Hojas	
6	44	←Para hojas de 0 a 4
6	9	← Para hojas de 5 a 9
7	01112334444	
7	555555666778899	
8	001112223334	
8	6899	
9	0024	
9		
10		
10		
11		
11		
12	0	

LA ESTADÍSTICA EN LAS NOTICIAS

El crecimiento de la estadística

El reportero Richard Rothstein escribió en el *New York Times* que el estudio del álgebra, la trigonometría y la geometría en la escuela preparatoria "deja muy poco espacio para el estudio de la estadística y la probabilidad. Sin embargo, los estudiantes necesitan fundamentos sobre el análisis de datos". El reportero observó que el cálculo tiene un papel prominente en los estudios universitarios, aun cuando "sólo algunos trabajos, principalmente en áreas técnicas, realmente lo utilizan". Rothstein citó un estudio realizado por el profesor Clifford Konold, de la Universidad de Massachusetts, quien contó el número de desplegados de datos que aparecen en el *New York Times*. En los ejemplares de 1972, el doctor Konold encontró cuatro gráficas o tablas en cada una de las 10 ediciones semanales (sin incluir las secciones de deportes y negocios), pero en 1982 había ocho, en 1992 fueron 44 y "el próximo año, él (el doctor Konold) podría encontrar más de 100". El crecimiento de la estadística como una disciplina se fomenta, en parte, por el uso creciente de dichos desplegados de datos en los medios de comunicación.

Cuando hay necesidad de *reducir* el número de renglones, es posible condensar una gráfica de tallo y hojas al combinar los renglones adyacentes, tal como se indica en la siguiente ilustración. Note que insertamos un asterisco para separar los dígitos en las hojas asociadas con los números en cada tallo. Cada renglón en la gráfica condensada debe incluir exactamente un asterisco, de modo que la forma de la gráfica no se distorsione.

Gráfica condensada de tallo y hojas

Tallo	Hojas	
6-7	449*0111233444455555566677 8899	← 64, 64, 69, 70, . . . , 79
8-9	0011122233346899*0024	
10-11	*	
12-13	0*	← El valor es 120.

Gráficas de Pareto

La Federal Communications Commission (FCC) verifica la calidad del servicio telefónico en Estados Unidos. Algunas de las quejas en contra de las compañías telefónicas incluyen los cambios, es decir, se cambia de compañía al cliente sin su consentimiento, y el cobro forzoso de cargos no autorizados. Datos recientes de la FCC mostraron que las quejas en contra de las compañías telefónicas estadounidenses eran las siguientes: 4473 por tarifas y servicios, 1007 por marketing, 766 por llamadas internacionales, 614 por cargos de acceso, 534 por servicios de operadora, 12,478 por cambios sin consentimiento y 1214 por forzamiento. Si usted fuese reportero de un medio impreso, ¿cómo presentaría dicha información? La simple escritura de oraciones con datos numéricos no llevaría a una verdadera comprensión. Un mejor método consiste en utilizar una gráfica conveniente; en este caso, la gráfica de Pareto se adecuaría muy bien.

Una **gráfica de Pareto** es una gráfica de barras para datos cualitativos, donde las barras se ordenan de acuerdo con las frecuencias. Al igual que en los histogramas, las escalas verticales de las gráficas de Pareto representan frecuencias o frecuencias relativas. La barra más alta se coloca a la izquierda y las más pequeñas hacia la derecha. Al ordenar las barras por frecuencias, la gráfica enfoca la atención en las categorías más importantes. La figura 2-6 es una gráfica de Pareto que muestra con claridad que el cambio sin consentimiento es, por mucho, el asunto más grave de las quejas de los clientes respecto de las empresas telefónicas.

Gráficas circulares

Las gráficas circulares también se utilizan para visualizar datos cualitativos. La figura 2-7 es un ejemplo de una **gráfica circular**, que presenta datos cualitativos como si fueran rebanadas de un pastel. La figura 2-7 representa los mismos datos de la figura 2-6. Para construir una gráfica circular, se separa el círculo en las proporciones que se adecuan mejor. La categoría de quejas por cambio sin consentimiento representan un 59% del total, de manera que la porción que representa el cambio sin consentimiento debe abarcar el 59% del total (con un ángulo central de $0.59 \times 360° = 212°$).

La gráfica de Pareto (figura 2-6) y la gráfica circular (figura 2-7) presentan los mismos datos en formas diferentes, pero una comparación probablemente demuestre que la gráfica de Pareto es mejor para resaltar los tamaños relativos de los distintos componentes, lo cual explica por qué muchas compañías, como Boeing Aircraft, a menudo utilizan las gráficas de Pareto.

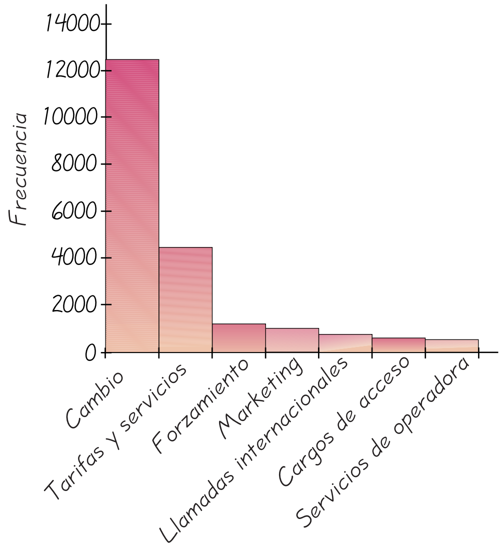


FIGURA 2-6 Gráfica de Pareto de quejas en contra de las compañías telefónicas

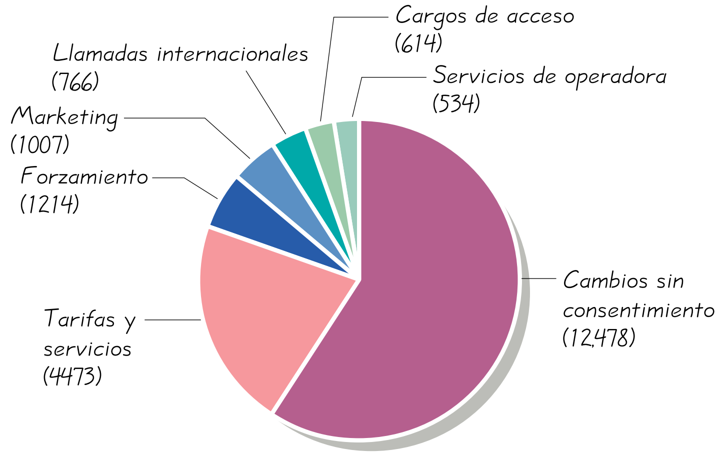
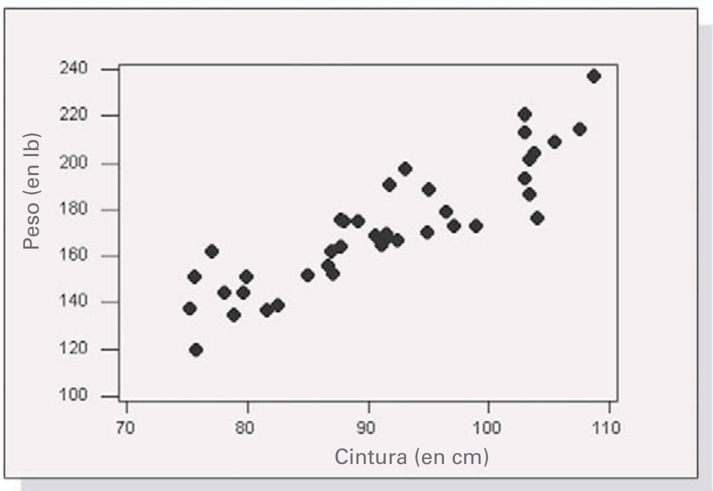


FIGURA 2-7 Gráfica circular de quejas en contra de las compañías telefónicas

Diagramas de dispersión

Un **diagrama de dispersión** es una gráfica de datos apareados (x, y), con un eje x horizontal y un eje y vertical. Los datos se aparean de tal forma que cada valor de un conjunto de datos corresponde a un valor de un segundo conjunto de datos. Para elaborar un diagrama de dispersión manualmente, construya un eje horizontal para los valores de la primera variable y un eje vertical para los valores de la segunda variable, y después grafique los puntos. El patrón de los puntos graficados suele ser útil para determinar si hay alguna relación entre las dos variables. (Este aspecto se estudia a profundidad en el tema de la correlación, en la sección 9-2). Con los datos del peso (en libras) y la circunferencia de la cintura (en cm) de los varones del conjunto de datos 1 del Apéndice B, utilizamos Minitab para generar el diagrama de dispersión que aparece a continuación. Con base en dicha gráfica, parece haber una relación entre el peso y la circunferencia de la cintura, tal como lo muestra el patrón de puntos.

Minitab



Florence Nightingale

A Florence Nightingale (1820-1910) se le reconoce como la fundadora de la profesión de enfermería, aunque también salvó miles de vidas usando la estadística. Cuando encontraba un hospital insalubre y con desabasto, mejoraba dichas condiciones y después utilizaba la estadística para convencer a otros de la necesidad de una reforma médica amplia. Ella diseñó gráficas originales para ilustrar que, durante la guerra de Crimea, murieron más soldados a consecuencia de las condiciones insalubres que en combate. Florence Nightingale fue pionera en el uso de la estadística social y de las técnicas gráficas.

Gráficas de series de tiempo

Los **datos de series de tiempo** son aquellos que se reúnen en diferentes momentos. Por ejemplo, la figura 2-8 muestra el número de pantallas de autocinemas existentes durante un periodo de 14 años (con base en datos de la National Association of Theater Owners). Vemos que durante este tiempo hay una clara tendencia de valores decrecientes. Lo que alguna vez fue parte importante de Estados Unidos, en especial para el autor, está en decadencia. Afortunadamente, la tasa de disminución parece ser menor que a finales de la década de 1980. Con frecuencia es sumamente importante conocer los cambios en los valores de una población a través del tiempo. Muchas compañías cayeron en la bancarrota porque no verificaban la calidad de sus bienes o servicios; además, de manera incorrecta, creían estar tratando con datos estables. No se dieron cuenta de que sus productos se volvían defectuosos conforme cambiaban importantes características de la población. El capítulo 13 introduce las *gráficas de control*, que son herramientas eficaces para verificar datos de series de tiempo.

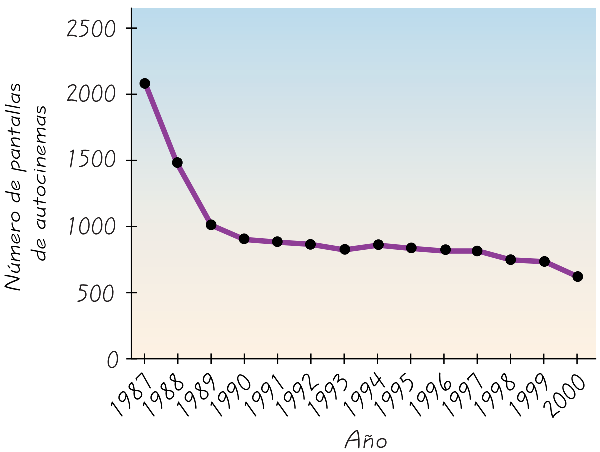
Otras gráficas

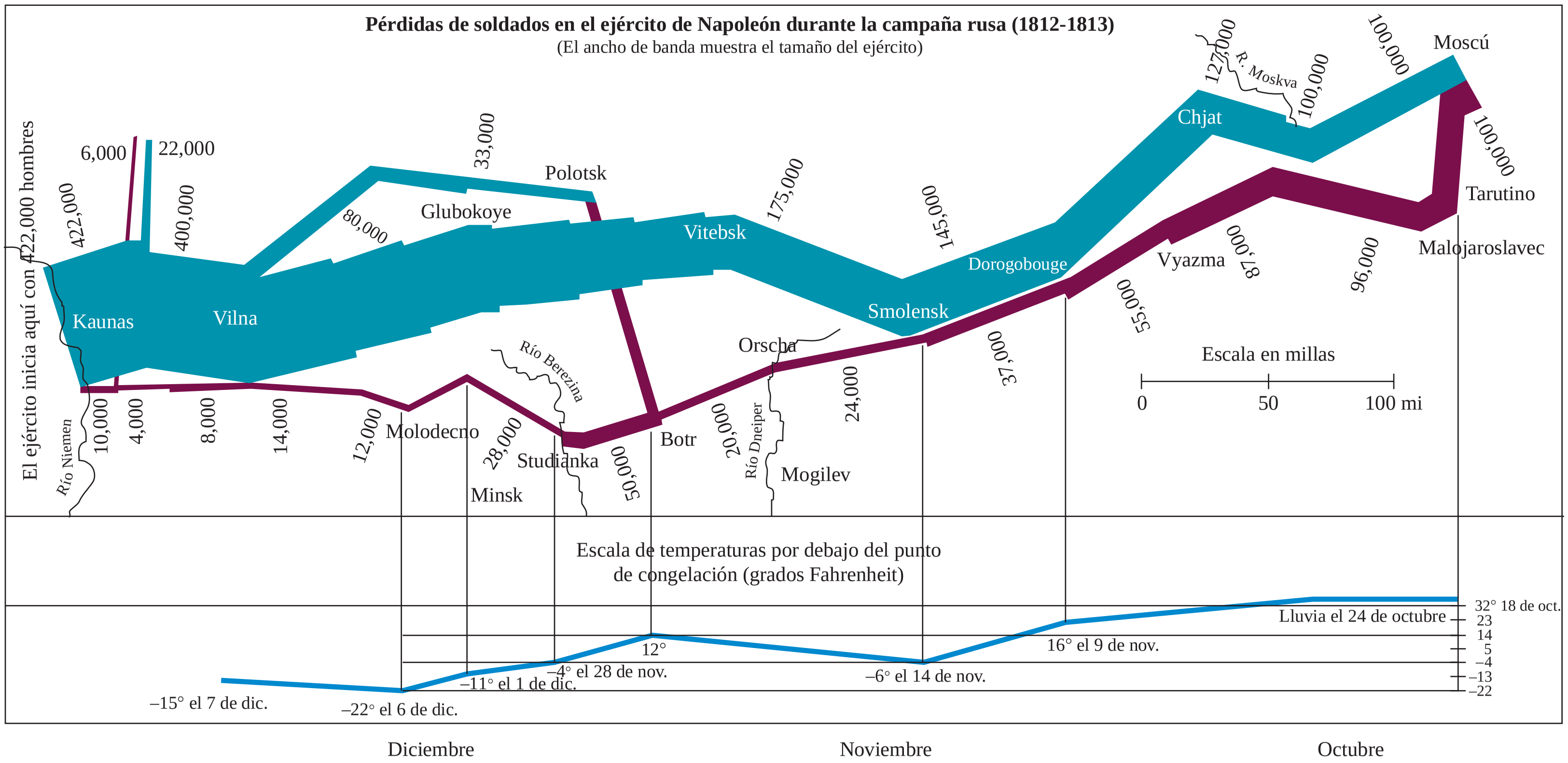
Además de las presentaciones gráficas descritas, hay muchas otras que pueden utilizarse para representar datos de manera llamativa y efectiva. En la sección 2-7 presentaremos las gráficas de cuadro, que son muy útiles para revelar la distribución de los datos. Los pictogramas representan datos con el uso de imágenes de objetos como soldados, tanques, aviones, monedas o bolsas de dinero.

La figura que aparece en la página 53 se ha descrito quizá como "la mejor gráfica estadística que se haya dibujado jamás". Esta figura incluye seis variables diferentes con respecto a la marcha del ejército de Napoleón hacia Moscú en 1812-1813. La banda gruesa a la izquierda representa el tamaño del ejército cuando inició la invasión a Rusia, desde Polonia. La banda inferior muestra su tamaño durante la retirada, con las temperaturas y fechas correspondientes. Aunque Charles Joseph Minard la elaboró en 1861, esta gráfica es ingeniosa, incluso desde la perspectiva actual.

Otra gráfica notable, de importancia histórica, es la que elaboró la enfermera más famosa del mundo, Florence Nightingale. Esta gráfica, que aparece en la figura 2-9 de la página 54, es particularmente interesante porque salvó vidas cuando Nightingale la utilizó para convencer a los oficiales británicos de que los hospitales militares necesitaban mejorar sus condiciones sanitarias, sus tratamientos y su abastecimiento. Su dibujo se asemeja a una gráfica circular, pese a que todos los ángulos centrales

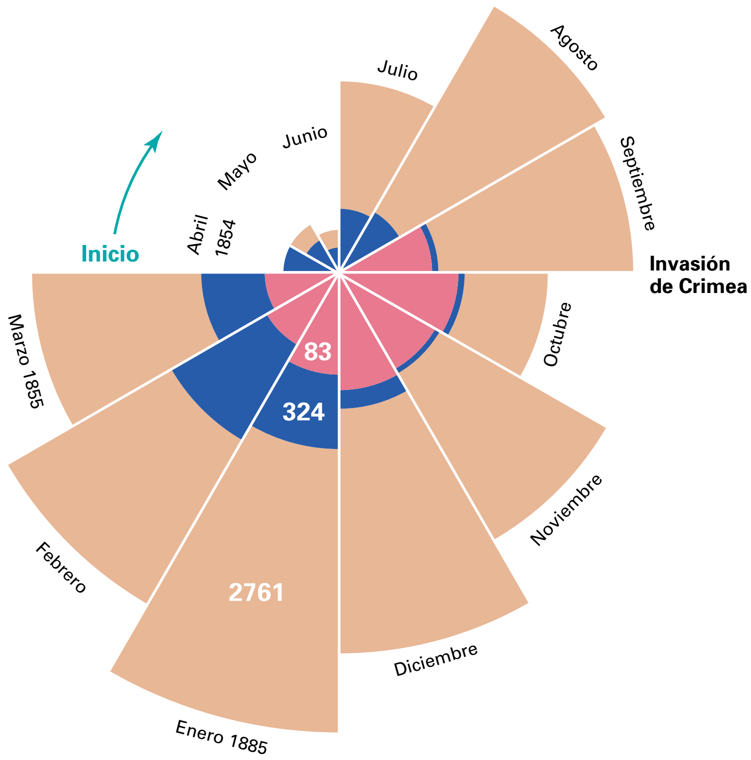
FIGURA 2-8 Datos de series de tiempo: número de pantallas de autocinemas





Créditos: Edward R. Tufte, *The Visual Display of Quantitative Information* (Cheshire, CT: Graphics Press, 1983). Se reproduce con autorización.

FIGURA 2-9 **Muertes en los hospitales militares británicos durante la guerra de Crimea**



son iguales y se usan radios diferentes para mostrar los cambios en el número de muertes mensuales. Las regiones externas de la figura 2-9 representan las muertes por enfermedades que pudieron prevenirse; las regiones internas representan muertes por heridas y las regiones centrales, muertes por otras causas.

Conclusión

La eficacia de la gráfica de Florence Nightingale ilustra muy bien el siguiente punto importante: una gráfica no es, en sí misma, un resultado final, es una herramienta para describir, explorar y comparar datos, que consideramos como sigue:

Descripción de datos: En un histograma, por ejemplo, se toman en cuenta el centro, la variación, la distribución y los datos distantes (CVDDT, sin el último elemento del tiempo). ¿Cuál es el valor aproximado del centro de la distribución y cuál es el rango aproximado de valores? Considere la forma completa de la distribución. ¿Están los valores distribuidos de manera uniforme? ¿La distribución está sesgada (ladeada) hacia la izquierda o hacia la derecha? ¿Tiene la distribución un pico a la mitad? Identifique cualquier valor extremo y cualquiera otra característica notable.

Exploración de datos: Buscamos características de la gráfica que revelen rasgos interesantes y/o útiles del conjunto de datos. Por ejemplo, en la figura 2-9 observamos que morían más soldados por cuidados hospitalarios inadecuados que por heridas de batalla.

Comparación de datos: Construya gráficas similares que faciliten la comparación de conjuntos de datos. Por ejemplo, si usted grafica un polígono de frecuencias con los pesos de hombres y otro polígono de frecuencias con pesos de mujeres, sobre el mismo conjunto de ejes, el polígono de los hombres debe aparecer a la derecha del polígono de mujeres, mostrando así que los hombres tienen pesos mayores.

Utilizando la tecnología

Ahora existen poderosos programas de computación que son bastante efectivos para generar gráficas impresionantes. Este libro hace referencia frecuente al STATDISK, Minitab, Excel y a la calculadora TI-83 Plus, por lo que listamos las gráficas (que ya comentamos en esta sección) que es posible elaborar. (Para información a detalle sobre los procedimientos, véanse los manuales que complementan este libro).

STATDISK Puede generar histogramas y diagramas de dispersión.

Minitab Puede generar todas las gráficas incluidas en esta sección.

Excel Puede generar histogramas, polígonos de frecuencias, gráficas circulares y diagramas de dispersión.

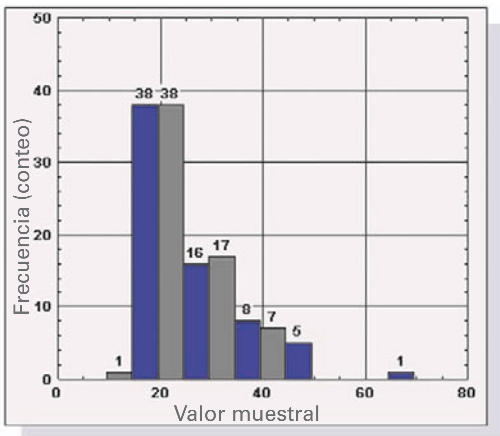
TI-83 Plus Puede generar histogramas y diagramas de dispersión.

2-3 Destrezas y conceptos básicos

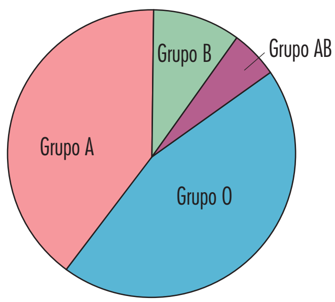
En los ejercicios 1 a 4, conteste las preguntas con respecto al histograma que se genera con STATDISK, el cual representa las edades de todos los polizontes del Queen Mary.

1. Centro ¿Cuál es el valor aproximado del centro? Es decir, ¿qué edad parece estar cerca del centro de todas las edades?
2. Variación ¿Cuáles son las edades más bajas y más altas posibles?
3. Porcentaje ¿Qué porcentaje de los 131 polizontes tenía menos de 30 años de edad?
4. Anchura de clase ¿Cuál es la anchura de clase?

STATDISK



En los ejercicios 5 y 6, remítase a la gráfica circular adjunta referente a los grupos sanguíneos de una muestra grande de personas (con base en datos del Greater New York Blood Program).



5. Interpretación de la gráfica circular ¿Cuál es el porcentaje aproximado de individuos con sangre tipo A? Suponiendo que la gráfica circular se base en una muestra de 500 personas, ¿aproximadamente cuántas de ellas tienen sangre tipo A?
6. Interpretación de la gráfica circular ¿Cuál es el porcentaje aproximado de personas con sangre tipo B? Suponiendo que la gráfica circular se base en una muestra de 500 personas, ¿aproximadamente cuántas de ellas tienen sangre tipo B?

Tabla del ejercicio 7

Edad	Estudiantes	Profesores
0–2	23	30
3–5	33	47
6–8	63	36
9–11	68	30
12–14	19	8
15–17	10	0
18–20	1	0
21–23	0	1

Tabla del ejercicio 8

Velocidad	Frecuencia
42–45	25
46–49	14
50–53	7
54–57	3
58–61	1

7. Automóviles de estudiantes/profesores Se obtuvieron muestras de automóviles de estudiantes y profesores en la universidad donde trabaja el autor. Sus edades (en años) se resumen en la distribución de frecuencias adyacente. Construya un histograma de frecuencias relativas para los automóviles de los estudiantes y otro histograma de frecuencias relativas para los automóviles de los profesores. Compare ambos. ¿Cuáles son las diferencias más notables?

8. Infracciones por exceso de velocidad La distribución de frecuencias adyacente describe las velocidades de conductores a quienes infraccionó la policía en la ciudad de Poughkeepsie. Estos conductores viajaban en una zona con límite de velocidad de 30 millas/hora en Creek Road, que atraviesa la universidad del autor. Construya un histograma correspondiente a la distribución de frecuencias. ¿Qué sugiere la distribución sobre el límite de velocidad establecido, comparándolo con el límite de velocidad señalado?

9. Osos El ejercicio 15, en la sección 2-2, se refiere al conjunto de datos 9 del Apéndice B. Use la distribución de frecuencias del peso de los osos (con 11 clases, iniciando con un límite de clase inferior de 0 y una anchura de clase de 50 lb), y construya el histograma correspondiente. ¿Cuál es el peso aproximado que se encuentra en el centro?

10. Temperaturas corporales El ejercicio 16, en la sección 2-2, se refiere al conjunto de datos 4 del Apéndice B. Con la distribución de frecuencias de las temperaturas corporales de medianoche del segundo día (con ocho clases, iniciando con el límite de clase inferior de 96.5 y una anchura de clase de 0.4°F), elabore el histograma correspondiente. ¿Qué sugiere la distribución sobre la creencia común de que la temperatura corporal promedio es de 98.6°F? Si se selecciona a los sujetos de forma aleatoria, las temperaturas deberían tener una distribución aproximadamente normal. ¿Es así?

En los ejercicios 11 a 14, realice las comparaciones construyendo las gráficas que se indican.

11. Circunferencia de cabezas El ejercicio 17, en la sección 2-2, se refiere al conjunto de datos 3 del Apéndice B. Utilice la distribución de frecuencias de la circunferencia de la cabeza de los niños y la distribución de frecuencias de la circunferencia de la cabeza de las niñas (con las clases de 34.0-35.9, 36.0-37.9, etcétera.), y construya los dos polígonos de frecuencias correspondientes, utilizando el mismo conjunto de ejes. Compare los resultados y determine si parece haber una diferencia significativa entre los dos géneros.

12. Películas de dibujos animados para niños El ejercicio 18, en la sección 2.2, se refiere al conjunto de datos 7 del Apéndice B. Utilice la distribución de frecuencias de la duración de las películas de dibujos animados para niños que incluyen consumo de tabaco y la distribución de frecuencias de la duración de aquéllas que presentan escenas de consumo de alcohol (con clases de 0-99, 100-199, etcétera.), y construya los dos polígonos de frecuencias correspondientes, usando el mismo conjunto de ejes. Compare los resultados y determine si parece haber una diferencia significativa.

13. Corredores del maratón El ejercicio 19, en la sección 2-2, se refiere al conjunto de datos 8 del Apéndice B. Utilice la distribución de frecuencias relativas de las edades de varones y la distribución de frecuencias relativas de las edades de mujeres (con un límite de clase inferior de 19 y una anchura de clase de 10), y construya los histogramas de frecuencias relativas correspondientes. Compare los resultados y determine si parece haber diferencias notables entre los dos grupos.

14. Coca Cola regular y Coca Cola dietética Remítase al conjunto de datos 17 del Apéndice B y utilice los pesos de la Coca Cola regular y de la Coca Cola dietética. Con las clases de 0.7750-0.7799, 0.7800-0.7849, . . . , 0.8250-0.8299, construya los dos polígonos de frecuencias sobre los mismos ejes. Después, compare los resultados y determine si parece haber una diferencia significativa. ¿Cuál sería una posible explicación de la diferencia?

En los ejercicios 15 y 16, liste los datos originales que se representan con las gráficas de tallo y hojas.

15.

Tallo (decenas)	Hojas (unidades)
20	0005
21	69999
22	2233333
23	
24	1177

16.

Tallo (centenas)	Hojas (decenas y unidades)
50	12 12 12 55
51	
52	00 00 00 00
53	27 27 35
54	72

En los ejercicios 17 y 18, construya la gráfica de puntos con los datos que se representan en la gráfica de tallo y hojas del ejercicio dado.

17. Ejercicio 15

18. Ejercicio 16

En los ejercicios 19 y 20, elabore las gráficas de tallo y hojas para los conjuntos de datos que se indican, que se encuentran en el Apéndice B.

19. Osos La longitud (en pulgadas) de los osos en el conjunto de datos 9. (*Sugerencia*: Primero redondee las longitudes hacia el entero más cercano).

20. Plástico Los pesos (en libras) del plástico que desechan 62 amas de casa: remítase al conjunto de datos 23, e inicie redondeando los pesos hacia el decimal más cercano. (Utilice una gráfica expandida de tallo y hojas, con aproximadamente 11 renglones).

21. Empleos Se realiza un estudio para determinar la manera en que las personas obtienen empleo. La tabla incluye datos de 400 sujetos que se seleccionaron aleatoriamente. Los datos se basan en resultados del National Center for Carrer Strategies. Construya la gráfica de Pareto correspondiente a tales datos. Si alguien deseara obtener un empleo, ¿cuál parece ser el método más efectivo?

Fuentes de empleo de sujetos que se encuestaron	Frecuencia
Anuncios clasificados	56
Empresas que buscan ejecutivos	44
Contactos interpersonales	280
Envíos por correo	20

22. Empleos Remítase a los datos del ejercicio 21 y construya una gráfica circular. Compárela con la gráfica de Pareto. ¿Podría determinar cuál gráfica es más efectiva para mostrar la importancia relativa de las fuentes de empleo?

23. Descarrilamiento de trenes Un análisis del descarrilamiento de trenes mostró que 23 de éstos fueron causados por vías en mal estado, nueve por fallas en el equipo, 12 por errores humanos y seis por otras causas (con base en datos de la Federal Railroad Administration). Construya una gráfica circular que represente tales datos.

24. Descarrilamiento de trenes Remítase a los datos del ejercicio 23 y elabore una gráfica de Pareto. Compare dicha gráfica con la gráfica circular. ¿Podría determinar cuál de las gráficas es más efectiva para mostrar la importancia relativa de las causas de los descarrilamientos de trenes?

En los ejercicios 25 y 26, utilice los datos apareados del apéndice B para construir un diagrama de dispersión.

25. Alquitrán/ CO Para el conjunto de datos 5, ubique el alquitrán en la escala horizontal y el monóxido de carbono en la escala vertical. Determine si parece haber una relación entre el alquitrán y el monóxido de carbono. Si es así, describa dicha relación.

26. Cuello/peso de osos Para el conjunto de datos 9, ubique las medidas del cuello en la escala horizontal y los pesos de los osos en la escala vertical. Con base en los resultados, ¿cuál es la relación existente entre el tamaño del cuello y el peso de los osos?

En los ejercicios 27 y 28, use los datos del Apéndice B para construir una gráfica de series de tiempo.

27. Inversiones en acciones Para el conjunto de datos 25, utilice los valores altos del Dow Jones Industrial Average (DJIA) para construir una gráfica de series de tiempo; después, determine si parece haber alguna tendencia. ¿Cómo podría un inversionista beneficiarse de esta tendencia?

28. Muertes en vehículos automotores En el conjunto de datos 25, utilice los datos de las muertes en vehículos automotores en Estados Unidos para construir una gráfica de series de tiempo; después, determine si parece haber alguna tendencia. Si es así, ofrezca una posible explicación.

En los ejercicios 29 a 32, remítase a la figura de la página 53, que describe la campaña de Napoleón de 1812 hacia Moscú y su retirada. La banda gruesa a la izquierda representa el tamaño del ejército cuando comenzó a invadir Rusia desde Polonia; la banda inferior describe la retirada de Napoleón.

29. Calcule el porcentaje de hombres que sobrevivieron toda la campaña.

30. Calcule el número de hombres y el porcentaje de hombres que murieron durante el cruce del río Berezina.

31. ¿Cuántos hombres murieron durante la retirada de Moscú, en el tiempo cuando la temperatura bajó de 16°F hasta -6°F?

32. De los hombres que lograron llegar a Moscú, ¿cuántos murieron en el viaje de regreso entre Moscú y Botr? (Observe que 33,000 hombres no fueron a Moscú, pero se unieron a los hombres que regresaban).

2-3 Más allá de lo básico

33. **a.** Remítase al conjunto de datos 20 del Apéndice B y elabore un histograma con las cargas axiales de las latas que tienen un grosor de 0.0111 pulgadas. El conjunto de datos incluye un dato distante de 504 lb. (Un dato distante es un valor que aparece muy lejos de los demás valores).
b. Repita el inciso *a*) después de excluir el dato distante de 504 lb.
c. ¿Qué efecto produce un dato distante en la forma del histograma?

34. Los Óscares En el artículo "Ages of Oscar-winning Best Actors and Actresses" (revista *Mathematics Teachers*), escrito por Richard Brown y Gretchen Davis, se utilizan gráficas de tallo y hojas para comparar las edades de los actores y las actrices en el momento que ganaron un Óscar. A continuación, se presentan los resultados de ganadores recientes, para cada categoría.

Actores:	32	37	36	32	51	53	33	61	35	45	55	39
	76	37	42	40	32	60	38	56	48	48	40	43
	62	43	42	44	41	56	39	46	31	47	45	60
	46	40	36									
Actrices:	50	44	35	80	26	28	41	21	61	38	49	33
	74	30	33	41	31	35	41	42	37	26	34	34
	35	26	61	60	34	24	30	37	31	27	39	34
	26	25	33									

Edades de los actores (unidades)	Tallo (decenas)	Edades de las actrices (unidades)
	2	
72	3	
	4	4
	5	0
	6	
	7	
	8	

a. Construya una gráfica de tallo y hojas, espalda con espalda, con los datos. Las primeras dos edades de cada grupo se insertaron al margen.

b. Utilice los resultados del inciso *a*), compare los dos conjuntos de datos y explique cualquier diferencia.

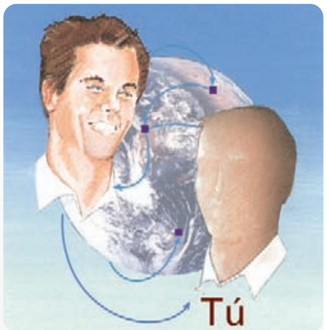
2-4 Medidas de tendencia central

Recuerde que el principal objetivo de este capítulo es lograr manejar las herramientas básicas para medir y describir diferentes características de un conjunto de datos. En la sección 2-1 observamos que, cuando describimos, exploramos y comparamos conjuntos de datos, las siguientes características suelen ser extremadamente importantes: centro, variación, distribución, datos distantes, cambios a través del tiempo. Las siglas CVDDT ("Cuidado con los Virus que Destruyen Datos y Trabajo") son útiles para recordar dichas características. En las secciones 2-2 y 2-3 señalamos que las distribuciones de frecuencias y las gráficas, así como los histogramas, sirven para investigar la distribución. En esta sección trataremos las características del *centro*.

Definición

Medida de tendencia central: valor que se encuentra en el centro o a la mitad de un conjunto de datos.

Hay muchas formas distintas de determinar el centro; por lo tanto, tenemos diferentes definiciones de las medidas de tendencia central, incluyendo media, mediana, moda y mitad del rango. Comenzaremos con la media.



Seis grados de separación

Los psicólogos sociales, historiadores, científicos políticos y especialistas en comunicaciones se interesan en el "problema del mundo pequeño": dadas dos personas cualesquiera en el mundo, ¿cuántos vínculos intermedios son necesarios para conectarlas? El psicólogo social Stanley Milgram realizó un experimento donde algunos sujetos intentaron ponerse en contacto con otras personas en específico, enviando por correo un archivo de información a un conocido que ellos pensaban estaba más cerca de la persona que buscaban. De las 160 cadenas de este tipo que se iniciaron, sólo 44 se completaron. El número de conocidos intermediarios varió de entre 2 y 10, con una mediana de 5 (o "seis grados de separación"). El experimento fue criticado por incluir sujetos muy sociales y por no hacer ajustes a las muchas conexiones perdidas de personas con ingresos más bajos. Otro estudio matemático mostró que si las cadenas perdidas se hubiesen completado, la mediana sería ligeramente mayor que 5.

Media

La media (aritmética) generalmente es la más importante de todas las medidas numéricas utilizadas para describir datos; constituye lo que la mayoría de la gente denomina *promedio*.

Definición

Media aritmética (de un conjunto de puntajes): medida de tendencia central que se obtiene sumando los puntajes y dividendo el total entre el número de puntajes. Tal medida de tendencia central se utilizará de manera frecuente a lo largo del libro; además, nos referiremos a ella simplemente como la **media**.

Esta definición se expresa como la fórmula 2-1, que utiliza la letra griega Σ (sigma mayúscula) para indicar que los valores de los datos deben sumarse. Esto es, Σx representa la sumatoria de todos los valores de los datos. El símbolo n denota el **tamaño de la muestra**, que es el número de puntajes en el conjunto de datos.

Fórmula 2-1
$$\text{media} = \frac{\Sigma x}{n}$$

La media se denota como $\bar{x}$ (se denomina "x barra"), si el conjunto de datos es una *muestra* de una población más grande; si se utilizan todos los puntajes de la población, entonces la media se simboliza con μ (mu minúscula). (Los estadísticos de una muestra generalmente se representan con letras inglesas, tales como $\bar{x}$, y los parámetros de la población con letras griegas, tales como μ).

Notación

Σ	denota la *sumatoria* de un conjunto de valores.
x	es la *variable* que suele utilizarse para representar los valores de datos individuales.
n	representa el *número de valores* de una *muestra*.
N	representa el *número de valores* de una *población*.
$\bar{x} = \frac{\Sigma x}{n}$	es la media de un conjunto de valores *muestrales*.
$\mu = \frac{\Sigma x}{N}$	es la media de todos los valores de una *población*.

EJEMPLO Verificación del plomo en el aire A continuación se presentan cantidades de plomo medidas (en microgramos por metro cúbico o $\mu g/m^3$) en el aire. La Environmental Protection Agency estableció un estándar de calidad del aire respecto del plomo: 1.5 $\mu g/m^3$. Las mediciones que se presentan más adelante se registraron en el edificio 5 del World Trade Center en distintos días, inmediatamente después de la destrucción causada por los ataques terroristas del 11 de septiembre de 2001. Después del colapso de los dos edificios del World Trade Center, surgió una gran preocupación respecto de la calidad del aire. Calcule la media de esta muestra de niveles de plomo en el aire.

5.40 1.10 0.42 0.73 0.48 1.10

SOLUCIÓN La media se calcula empleando la fórmula 2-1. Primero se suman los puntajes y después se dividen entre el número de ellos:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{5.40 + 1.10 + 0.42 + 0.73 + 0.48 + 1.10}{6} = \frac{9.23}{6} = 1.538$$

La media del nivel de plomo es 1.538 $\mu g/m^3$. Además del valor de la media, también es notable que el conjunto de datos incluye un valor (5.40), que está muy distante de los demás. Sería importante investigar un "dato distante" como éste. En tal caso, el nivel de plomo de 5.40 $\mu g/m^3$ se midió un día después del colapso de las torres del World Trade Center y los niveles de polvo y humo ofrecieron una explicación razonable para un valor tan extremo.

Una desventaja de la media es su sensibilidad a cada valor, de modo que un puntaje excepcional puede afectarla de manera drástica. La mediana resuelve, en gran parte, esa desventaja.

Mediana

Definición

Mediana (de un conjunto de datos): medida de tendencia central que implica el valor que está en medio, cuando los valores originales de los datos se presentan en orden de magnitud creciente (o decreciente). La mediana suele denotarse con $\tilde{x}$ (se pronuncia "*x* con tilde").

Para calcular la mediana, primero clasifique los valores (acomódelos en orden), luego siga uno de estos dos procedimientos:

1. Si el número de valores es impar, la mediana es el número que se localiza exactamente a la mitad de la lista.
2. Si el número de valores es par, la mediana se obtiene calculando la media de los dos números que están a la mitad.

La figura 2-10 demuestra este procedimiento para el cálculo de la mediana.

EJEMPLO Verificación del plomo en el aire A continuación se presentan cantidades de plomo medidas (en $\mu g/m^3$) en el aire. Calcule la mediana de esta muestra.

5.40 1.10 0.42 0.73 0.48 1.10

SOLUCIÓN Primero ordene los valores:

0.42 0.48 0.73 1.10 1.10 5.40

continúa

Paradoja del tamaño del grupo escolar

Hay al menos dos formas de obtener la media de un grupo escolar, y pueden dar resultados muy distintos. En una universidad, si consideramos las cantidades de estudiantes en 737 grupos, obtenemos una media de 40 estudiantes. Pero si compilamos una lista de los tamaños de grupo para cada estudiante y utilizamos dicha lista, obtendríamos una media de grupo de 147. Esta gran discrepancia es por el hecho de que hay muchos estudiantes en los grupos grandes, mientras que en los grupos pequeños hay muy pocos. Sin modificar el número de grupos ni de profesores, reduciríamos el tamaño medio del grupo escolar que los estudiantes experimentan, haciendo que todos los grupos tengan aproximadamente el mismo tamaño. Lo anterior también incrementaría la asistencia, que es mayor en los grupos escolares pequeños.

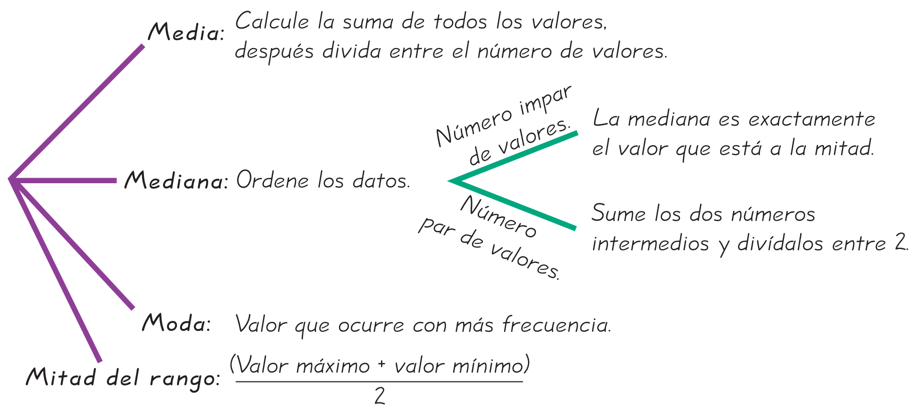


FIGURA 2-10 Procedimientos para calcular las medidas de tendencia central

Puesto que el número de valores es par (6), la mediana se obtiene calculando la media de los dos valores intermedios 0.73 y 1.10.

$$\text{Mediana} = \frac{0.73 + 1.10}{2} = \frac{1.83}{2} = 0.915$$

Como el número de valores es par (6), la mediana es el número que se encuentra exactamente a la mitad de la lista ordenada; por lo tanto, la mediana es 0.915 $\mu g/m^3$. Note que la mediana es muy diferente de la media de 1.538 $\mu g/m^3$, que se obtuvo del mismo conjunto de datos muestrales del ejemplo anterior. La razón de esa gran discrepancia es el efecto que el puntaje 5.40 tuvo en la media. Si este valor extremo se redujera a 1.20, la media caería de 1.538 $\mu g/m^3$ hasta 0.838 $\mu g/m^3$, pero la mediana no cambiaría.

EJEMPLO Verificación del plomo en el aire Repita el ejemplo anterior, después de incluir la medición de 0.66 $\mu g/m^3$, que se registró otro día. Es decir, calcule la mediana de estas mediciones del plomo:

5.40 1.10 0.42 0.73 0.48 1.10 0.66

SOLUCIÓN Primero ordene los valores

0.42 0.48 0.66 0.73 1.10 1.10 5.40

Puesto que el número de valores es impar (7), la mediana es exactamente el valor a la mitad de la lista ordenada: 0.73 $\mu g/m^3$.

Después de estudiar los ejemplos anteriores, debe quedar claro el procedimiento para obtener la mediana; además, que la media se ve afectada de manera drástica por valores extremos, mientras que la mediana no. Puesto que la mediana no es tan sensible a los valores extremos, con frecuencia se utiliza para conjuntos de datos que tienen un número relativamente pequeño de datos distantes. Por ejemplo, recientemente la oficina de censos de Estados Unidos reportó que la *mediana* del ingreso familiar es de 36,078 dólares anuales. Se usó la mediana ya que existen pocas familias con ingresos realmente altos.

Moda

Definiciones

Moda (de un conjunto de datos, que suele denotarse como M): valor que ocurre con mayor frecuencia.

- Cuando dos valores ocurren con la misma frecuencia y ésta es la más alta, ambos valores son modas, por lo que el conjunto de datos es **bimodal.**
- Cuando más de dos valores ocurren con la misma frecuencia y ésta es la más alta, todos los valores son modas, por lo que el conjunto de datos es **multimodal.**
- Cuando ningún valor se repite, se dice que no hay moda.

EJEMPLO Calcule las modas de los siguientes conjuntos de datos.

a. 5.40 1.10 0.42 0.73 0.48 1.10

b. 27 27 27 55 55 55 88 88 99

c. 1 2 3 6 7 8 9 10

SOLUCIÓN

a. El número 1.10 es la moda, ya que es el valor que ocurre con mayor frecuencia.

b. Los números 27 y 55 son modas, ya que ambos ocurren con la frecuencia más alta. Este conjunto de datos es bimodal, porque tiene dos modas.

c. No hay moda, ya que ningún valor se repite.

En realidad, la moda no se utiliza mucho con datos numéricos. Sin embargo, entre las distintas medidas de tendencia central que consideramos, la moda es la única que puede usarse con datos de nivel nominal de medición. (Recuerde que el nivel nominal de medición se aplica a datos que consisten únicamente en nombres, etiquetas o categorías). Por ejemplo, una encuesta que se aplicó a estudiantes universitarios mostró que el 84% tiene aparato de televisión; el 76%, videocasetera; el 60%, reproductor de discos compactos portátil; el 39%, sistema de videojuegos y el 35%, reproductor de DVD (con base en datos del National Center for Education Statistics). En tanto que el televisor es el aparato más frecuente, es posible afirmar que la moda es el televisor. No podemos calcular una media o mediana para datos como éstos, a nivel nominal.

Mitad del rango

Definición

Mitad del rango: medida de tendencia central que constituye el valor que está a medio camino, entre el puntaje más alto y el más bajo, en el conjunto original de datos. Se calcula sumando el valor máximo con el mínimo y luego dividiendo dicha suma entre 2, como en la siguiente fórmula.

$$\text{mitad del rango} = \frac{(\text{valor máximo} + \text{valor mínimo})}{2}$$

LA ESTADÍSTICA EN LAS NOTICIAS

Un hombre promedio

La revista *Men´s Health* publicó estadísticas que describen al “hombre promedio”, que tiene 34.4 años de edad, pesa 175 libras, mide cerca de 5 pies 10 pulgadas y se llama Mike Smith. La edad, el peso y la estatura son valores medios, pero el nombre de Mike Smith es la moda que corresponde al nombre y apellido más comunes. Otra estadística notable es la siguiente: el hombre promedio duerme aproximadamente 6.9 horas por noche, bebe cerca de 3.3 tasas de café al día y consume 1.2 bebidas alcohólicas diariamente; además, gana alrededor de 36,100 dólares anuales, debe 2,563 dólares en las tarjetas de crédito y tiene 3,100 dólares ahorrados en el banco.

LA ESTADÍSTICA EN LAS NOTICIAS

Maniquís ≠ realidad

La revista *Health* comparó las medidas de los maniquís con las medidas de las mujeres. Los siguientes resultados se reportaron como "promedios", que tal vez representan medias. Estatura de los maniquíes: 6 pies; estatura de las mujeres: 5 pies 4 pulgadas. Cintura de los maniquíes: 23 pulgadas; cintura de las mujeres: 29 pulgadas. Tamaño de la cadera de los maniquíes: 34 pulgadas; tamaño de la cadera de las mujeres: 40 pulgadas. Talla de vestido de los maniquíes: 6; talla de vestido de las mujeres: 11. Cuando se comparan las medias es evidente que los maniquíes y las mujeres reales son muy diferentes.

La mitad del rango se utiliza en pocas ocasiones. Puesto que sólo utiliza los valores máximo y mínimo es demasiado sensible a dichos extremos. Sin embargo, la mitad del rango posee tres características positivas: **1.** es fácil de calcular; **2.** ayuda a reforzar el hecho importante de que existen diferentes formas para definir el centro de un conjunto de datos; **3.** en ocasiones se utiliza de manera incorrecta como si fuese la mediana, de manera que es posible disminuir la confusión al definir con claridad la mitad del rango con respecto a la mediana.

EJEMPLO Verificación del plomo en el aire A continuación se presentan medidas de las cantidades de plomo ($\mu g/m^3$) en el aire, en el lugar donde estaba el World Trade Center, días después del 11 de septiembre del 2001. Calcule la mitad del rango para esta muestra:

5.40 1.10 0.42 0.73 0.48 1.10

SOLUCIÓN La mitad del rango se calcula de la siguiente manera:

$$\frac{(\text{valor máximo} + \text{valor mínimo})}{2} = \frac{(5.40 + 0.42)}{2} = 2.910$$

La mitad del rango es 2.910 $\mu g/m^3$.

Desafortunadamente, el término *promedio* en ocasiones se utiliza para cualquier medida de tendencia central y, en ocasiones, para implicar la media. Por esta ambigüedad, no debemos usar el término *promedio* cuando nos referimos a una medida de tendencia central en particular. En su lugar, habrá que aplicar el término específico, tal como media, mediana, moda o mitad del rango. Cuando nos encontremos un valor reportado como *promedio*, tendremos que saber que el valor puede ser el resultado de cualquiera de las distintas definiciones.

Con la idea de describir, explorar y comparar datos, incluimos la tabla 2-8, que resume las distintas medidas de tendencia central para los niveles de cotinina que se presentan en la tabla 2-1, en el problema del capítulo. Recuerde que la cotinina es un metabolito de la nicotina, de modo que cuando el cuerpo absorbe la nicotina se produce la cotinina. Una comparación de las medidas de tendencia central sugiere que los niveles de cotinina son más altos en los fumadores. Además, los niveles de cotinina de los individuos que no fuman, pero están expuestos al humo del tabaco, son más altos que los de personas que tampoco fuman y no están expuestas al humo. Lo anterior sugiere que "los fumadores pasivos" sí se ven afectados. Se

Tabla 2-8 Comparación de los niveles de cotinina de fumadores, de no fumadores expuestos al humo ambiental del tabaco (HAT) y de no fumadores no expuestos al humo ambiental del tabaco (SHAT).

	Fumadores	HAT	SHAT
Media	172.5	60.6	16.4
Mediana	170.0	1.5	0.0
Moda	1 y 173	1	0
Mitad del rango	245.5	275.5	154.5

dispone de métodos para determinar si estas aparentes diferencias son estadísticamente significativas. Más adelante, consideraremos algunos de estos métodos.

Regla de redondeo

Una regla sencilla para redondear respuestas es la siguiente:

> **Aumente una posición decimal más a las que están presentes en el conjunto original de datos.**

Cuando aplique esta regla, redondee sólo la respuesta final y *no los valores intermedios que aparecen durante los cálculos*. Así, la media de 2, 3, 5, es 3.333333 . . . , que se redondea a 3.3. Como los valores originales son números enteros, redondeamos al décimo más cercano. Otro ejemplo sería la media de 80.4 y 80.6, que es igual a 80.50 (una posición decimal más de la que se empleó para los valores originales).

Media de una distribución de frecuencias

Cuando los datos se resumen en una distribución de frecuencias es probable que no conozcamos los valores exactos de una clase en particular. Para hacer que los cálculos sean posibles, pretendemos que todos los valores muestrales sean iguales a la marca de clase. Ya que cada marca de clase se repite un número de veces igual a la frecuencia de clase, la sumatoria de todos los valores muestrales es $\Sigma(f \cdot x)$, donde f denota la frecuencia y x representa la marca de clase. El número total de valores muestrales es la sumatoria de frecuencias Σf. La fórmula 2-2 se utiliza para calcular la media cuando los datos muestrales se resumen en una distribución de frecuencias. La fórmula 2-2 en realidad no es un concepto nuevo, sino una variación de la fórmula 2-1.

Primero multiplique cada frecuencia y marca de clase, después sume los productos
↓

Fórmula 2-2

$$\bar{x} = \frac{\Sigma(f \cdot x)}{\Sigma f}$$

(media de la distribución de frecuencias)

↑
sumatoria de las frecuencias

Por ejemplo, observe la tabla 2-9 en la siguiente página. Las primeras dos columnas son iguales a la distribución de frecuencias (tabla 2-2) de los niveles de cotinina de fumadores. La tabla 2-9 ilustra el procedimiento que se sigue para aplicar la fórmula 2-2 cuando se calcula la media de datos resumidos en una distribución de frecuencias. En realidad, por lo general se utilizan programas de cómputo o calculadoras, en lugar del cálculo manual. El resultado de la tabla 2-9 es $\bar{x} = 177.0$, aunque obtenemos $\bar{x} = 172.5$ si utilizamos la lista original con 40 valores. Recuerde, la distribución de frecuencias produce una aproximación de $\bar{x}$, ya que no se basa en la lista original exacta de valores muestrales.

Nadie en casa

Los encuestadores no pueden ignorar simplemente a quienes no estaban en casa cuando acudieron por primera vez. Una solución implica regresar varias veces hasta localizar a la persona. Alfred Politz y Willard Simmons describen una forma para compensar los resultados faltantes, sin tener que regresar varias veces. Sugieren ponderar los resultados con base en la frecuencia en que la gente no se encuentra en su casa. Por ejemplo, alguien que está en su casa sólo dos, de seis días a la semana, tendrá una probabilidad de 2/6 o 1/3 de estar allí en la primera visita. Cuando se localiza a dicha persona por primera vez, sus resultados se ponderan de modo que se cuenten tres veces, respecto de un sujeto que siempre está en su casa. Esta ponderación compensa a los demás individuos similares que permanecen en casa dos de seis días a la semana y que no respondieron cuando se les buscó por primera vez. Tan inteligente solución se presentó inicialmente en 1949.

Tabla 2-9 Cálculo de la media de una distribución de frecuencias

Nivel de cotinina	Frecuencia f	Marca de clase x	$f \cdot x$
0–99	11	49.5	544.5
100–199	12	149.5	1794.0
200–299	14	249.5	3493.0
300–399	1	349.5	349.5
400–499	2	449.5	899.0
Totales:	$\Sigma f = 40$		$\Sigma(f \cdot x) = 7080.0$

$$\bar{x} = \frac{\Sigma(f \cdot x)}{\Sigma f} = \frac{7080}{40} = 177.0$$

Media ponderada

En algunos casos los valores varían su grado de importancia, de modo que es posible que queramos acomodarlos de acuerdo con ello. Después, será posible proceder al cálculo de una **media ponderada**, que es una media que se obtiene asignando distintos pesos a los valores, tal como se muestra en la fórmula 2-3.

Fórmula 2-3 media ponderada: $\bar{x} = \dfrac{\Sigma(w \cdot x)}{\Sigma w}$

Por ejemplo, suponga que necesitamos una media de tres calificaciones de una prueba (85, 90, 75), donde la primera prueba cuenta el 20%, la segunda el 30% y la tercera el 50% de la calificación final. Podemos asignar pesos de 20, 30 y 50 a las calificaciones de la prueba y luego calcular la media aplicando la fórmula 2-3, como sigue:

$$\bar{x} = \frac{\Sigma(w \cdot x)}{\Sigma w}$$
$$= \frac{(20 \times 85) + (30 \times 90) + (50 \times 75)}{20 + 30 + 50} = \frac{8150}{100} = 81.5$$

Otro ejemplo son los promedios universitarios (que utilizan letras), que pueden calcularse si asignamos a cada calificación con letras el número adecuado de puntos (A = 4, B = 3, etcétera), y después asignamos a cada puntaje un peso igual al número de horas crédito. Nuevamente, se utiliza la fórmula 2-3 para calcular el promedio de calificaciones.

La mejor medida de tendencia central

Hasta ahora hemos considerado la media, mediana, moda y mitad del rango como medidas de tendencia central. ¿Cuál de ellas es la mejor? Desafortunadamente, no existe una respuesta única a esa pregunta, porque no hay criterios objetivos para determinar la medida más representativa para todos los conjuntos de datos. Las diferentes medidas de tendencia central ofrecen diversas ventajas y desventajas, algunas de las cuales se resumen en la tabla 2-10.

Tabla 2-10 Comparación de la media, mediana, moda y mitad del rango

Medida de tendencia central	Definición	¿Qué tan común es?	Existencia	¿Toma en cuenta cada valor?	¿Se ve afectada por valores extremos?	Ventajas y desventajas
Media	$\bar{x} = \frac{\Sigma x}{n}$	"promedio" más conocido	siempre existe	sí	sí	se usa a lo largo de este libro; funciona bien con muchos métodos estadísticos
Mediana	valor en medio	de uso común	siempre existe	no	no	suele ser una buena opción si hay algunos valores extremos
Moda	valor más frecuente	se usa en ocasiones	podría no existir; podría haber más de una	no	no	apropiada para datos en el nivel nominal
Mitad del rango	$\frac{(\text{máx} + \text{mín})}{2}$	poco usada	siempre existe	no	sí	muy sensible a los valores extremos

Comentarios generales:

- En el caso de una colección de datos que es aproximadamente simétrica con una moda, la media, la mediana, la moda y la mitad del rango tienden a ser iguales.
- En el caso de una colección de datos obviamente asimétrica, sería bueno reportar tanto la media como la mediana.
- La media es relativamente *confiable*. Es decir, cuando las muestras se extraen de la misma población, las medidas muestrales tienden a ser más consistentes que las demás medidas de tendencia central (consistentes en el sentido de que las medias muestrales, extraídas de la misma población, no varían tanto como las otras medidas de tendencia central).

Una ventaja importante de la media es que toma en cuenta cada valor; una desventaja notoria es que en ocasiones se ve afectada de manera drástica por unos cuantos valores extremos. Tal desventaja se supera usando una media recortada, como se describe en el ejercicio 21.

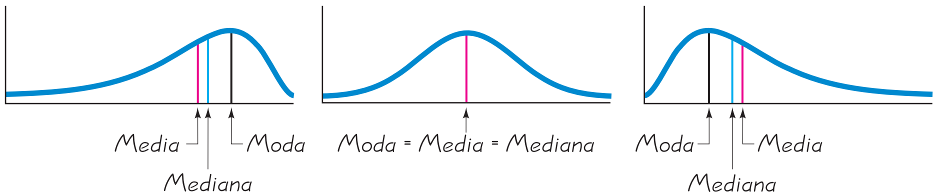
Sesgo

Una comparación de la media, la mediana y la moda puede revelar información acerca de la característica de sesgo, que se define a continuación y se ilustra en la figura 2-11.

Definición

Una distribución de datos está **sesgada** si no es simétrica y se extiende más hacia un lado que hacia el otro. (Una distribución de datos es **simétrica** si la mitad izquierda de su histograma es aproximadamente una imagen en espejo de su mitad derecha).

FIGURA 2-11 Sesgo



(a) Sesgada a la izquierda (sesgo negativo): la media y la mediana están a la izquierda de la moda.

(b) Simétrica (sesgo cero): la media, la mediana y la moda son iguales.

(c) Sesgada a la derecha (sesgo positivo): la media y la mediana están a la derecha de la moda.

Los datos **sesgados a la izquierda** (que también se denomina como *sesgo negativo*) poseen una cola izquierda más larga, en tanto que la media y la mediana se encuentran a la izquierda de la moda. Aunque no siempre es posible predecirlo, los datos sesgados a la izquierda suelen tener una media menor a la mediana, como sucede en la figura 2-11*a*). Los datos **sesgados a la derecha** (lo que también se denomina *sesgo positivo*) poseen una cola derecha más larga, mientras que la media y la mediana se encuentran a la derecha de la moda. Nuevamente, aunque no siempre es posible predecirlo, en los datos sesgados a la derecha, la media suele estar a la derecha de la mediana, como en la figura 2-11c).

Si examinamos el histograma de la figura 2-1, para los niveles de cotinina de fumadores, observaremos una gráfica sesgada hacia la derecha. En la práctica, muchas distribuciones de datos son simétricas y carecen de sesgo. Las distribuciones sesgadas hacia la derecha son más comunes que aquéllas sesgadas hacia la izquierda, porque con frecuencia es más fácil obtener valores excepcionalmente grandes que valores excepcionalmente pequeños. En el caso de los ingresos anuales, por ejemplo, es imposible obtener valores por debajo del límite inferior de cero, pero hay algunas personas que ganan millones de dólares en un año. Por lo tanto, los ingresos anuales tienden a estar sesgados hacia la derecha, como en la figura 2-11*c*).

Utilizando la tecnología

Los cálculos en esta sección son bastante sencillos, pero algunos de la siguiente sección requieren mayor esfuerzo. Muchos programas de cómputo le permiten introducir un conjunto de datos y utilizar una operación para obtener diversos estadísticos para muestras, que se engloban en la *estadística descriptiva*. (Véase la sección 2-6, donde se incluyen representaciones visuales de muestras que se obtienen con el STATDISK, Minitab, Excel y la calculadora TI-83 Plus.) A continuación, se incluyen algunos de los procedimientos para la obtención de dichas representaciones visuales.

STATDISK Elija **Data** del menú principal y utilice el **Sample Editor** para ingresar los datos. Presione **Copy** y luego presione **Data** nuevamente, pero ahora elija la opción **Descriptive Statistics**. Haga clic en **Paste**, para recuperar el conjunto de datos que usted alimentó. Ahora haga clic en **Evaluate** para obtener los estadísticos descriptivos, incluyendo la media, la mediana, la mitad del rango y otros que se discutirán en las próximas secciones.

Minitab Ingrese los datos en la columna que tiene el encabezado C1. Haga clic en **Stat**, seleccione **Basic Statistics** y luego **Descriptive Statistics.** Los resultados incluirán la media y la mediana, así como otros estadísticos.

Excel Ingrese los datos de la muestra en la columna A. Seleccione **Tools**, después **Data Analysis** y luego **Descriptive Statistics**; haga clic en **OK**. En el cuadro de diálogo, ingrese el rango de entrada de datos (tal como A1:A40 para 40 valores en la columna A), haga clic en **Summary Statistics** y después haga clic en **OK**. (Si Data Analysis no aparece en el menú Tools, deberá instalarlo haciendo clic en **Tools** y seleccionando **Add-Ins.**)

TI-83 Plus Primero ingrese los datos en la lista L1, presionando **STAT**, luego **Edit** y finalmente la tecla **ENTER**. Una vez que ingresó los datos, presione **STAT** y seleccione **CALC**, después **1-Var Stats** y finalmente **ENTER** dos veces. La representación visual incluirá la media, la mediana, el valor mínimo y el valor máximo. Utilice la flecha que va hacia abajo para ver los resultados que no aparecen en la primera representación visual.

2-4 Destrezas y conceptos básicos

En los ejercicios 1 a 8 calcule a) la media, b) la mediana, c) la moda y d) la mitad del rango, de los datos muestrales dados.

1. Consumo de tabaco en películas infantiles En el artículo "Tobacco and Alcohol Use in G-Rated Children's Animated Films", de Goldstein, Sobel y Newman (*Journal of the American Medical Association*, vol. 281, núm. 12), se registró la duración (en segundos) de escenas de películas animadas, de los Universal Studios, que muestran consumo de tabaco. A continuación se presentan los primeros seis valores, que están incluidos en el conjunto de datos 7 del Apéndice B. ¿Hay algún problema al incluir escenas de consumo de tabaco en películas infantiles de dibujos animados?

0	223	0	176	0	548

2. Harry Potter En un intento por medir el nivel de lectura de un libro, se obtuvieron los puntajes de la facilidad de lectura de Flesch Reading de 12 páginas, seleccionadas aleatoriamente, de la obra *Harry Potter y la piedra filosofal*, de J. K. Rowling. Dichos puntajes, que se incluyen en el conjunto de datos 14 del Apéndice B, se listan a continuación. En tanto que estos puntajes se basan en 12 páginas seleccionadas aleatoriamente, ¿será la media de esta muestra un estimado razonable del nivel medio de lectura de todo libro?

85.3	84.3	79.5	82.5	80.2	84.6
79.2	70.9	78.6	86.2	74.0	83.7

3. Cereal Un nutriólogo obtiene las cantidades de azúcar (en gramos) de un gramo de 16 cereales diferentes, incluyendo Cheerios, Corn Flakes, Fruit Loops, Trix y 12 más. Estos valores, que se incluyen en el conjunto de datos 16 del Apéndice B, se listan a continuación. ¿Será la media de estos valores un buen estimado de la cantidad media de azúcar que hay en cada gramo del cereal consumido por la población de todos los estadounidenses que lo comen? ¿Por qué?

0.03	0.24	0.30	0.47	0.43	0.07	0.47	0.13
0.44	0.39	0.48	0.17	0.13	0.09	0.45	0.43

4. Índice de masa corporal Como parte del examen nacional de salud (National Health Examination) en Estados Unidos, se mide el índice de masa corporal en una muestra aleatoria de mujeres. Algunos de los valores, que se incluyen en el conjunto de datos 1 del Apéndice B, se listan a continuación. ¿Estará la media de tal muestra razonablemente cerca de la media de 25.74, que es la media de las 40 mujeres incluidas en el conjunto de datos 1?

19.6	23.8	19.6	29.1	25.2	21.4	22.0	27.5
33.5	20.6	29.9	17.7	24.0	28.9	37.7	

5. Conductores alcoholizados Más adelante se incluyen las concentraciones de alcohol en la sangre de conductores que se vieron envueltos en accidentes fatales y que después fueron sentenciados a prisión (con base en datos del Departamento de Justicia de Estados Unidos). Si las leyes estatales actuales prohíben conducir con niveles por encima de 0.08 o 0.10, ¿están estos niveles significativamente por arriba del máximo permitido?

0.27	0.17	0.17	0.16	0.13	0.24	0.29	0.24
0.14	0.16	0.12	0.16	0.21	0.17	0.18	

6. Muertes en motocicletas A continuación se presentan las edades de motociclistas que se accidentaron mortalmente en accidentes de tránsito (con base en datos del Departamento del Transporte estadounidense). ¿Apoyan los resultados la creencia común de que una mayor proporción de conductores jóvenes se ven implicados en tales accidentes?

17	38	27	14	18	34	16	42	28
24	40	20	23	31	37	21	30	25

7. Tiempos de reacción El autor visitó el Museo de Ciencias Reuben H. Fleet en San Diego y repitió un experimento de tiempos de reacción. Se obtuvieron los siguientes tiempos (en centésimas de segundo). ¿Qué tan consistentes son estos resultados y de qué forma afecta la consistencia el uso de la media muestral, como un estimado de la media poblacional?

19	20	17	21	21	21	19	18	19	19
17	17	15	17	18	17	18	18	18	17

8. Tabletas de Bufferin A continuación se listan los pesos medidos (en miligramos) de una muestra de tabletas de aspirina Bufferin. ¿Cuál sería la consecuencia grave de tener pesos que varían tanto?

672.2	679.2	669.8	672.6	672.2	662.2
662.7	661.3	654.2	667.4	667.0	670.7

En los ejercicios 9 a 12 calcule media, mediana, moda y mitad del rango para cada una de las dos muestras; luego, compare los dos conjuntos de resultados.

9. Tiempos de espera de clientes A continuación se presentan los tiempos de espera (en minutos) de los clientes del Banco Jefferson Valley (donde todos los clientes forman una sola fila) y del Banco Providence (donde los clientes esperan en filas individuales, en tres ventanillas diferentes):

Jefferson Valley:	6.5	6.6	6.7	6.8	7.1	7.3	7.4	7.7	7.7	7.7
Providence:	4.2	5.4	5.8	6.2	6.7	7.7	7.7	8.5	9.3	10.0

Interprete los resultados y determine si hay una diferencia entre los dos conjuntos de datos, que no sea aparente cuando se comparen las medidas de tendencia central. Si es así, ¿cuál es?

10. Coca Cola regular/Coca Cola dietética Los siguientes son los pesos (en libras) de muestras del contenido de latas de Coca Cola regular y Coca Cola dietética:

Regular:	0.8192	0.8150	0.8163	0.8211	0.8181	0.8247
De dieta:	0.7773	0.7758	0.7896	0.7868	0.7844	0.7861

¿Parece haber una diferencia significativa entre los dos conjuntos de datos? ¿Cómo se explicaría una diferencia como ésa?

11. Mickey D *vs.* Jack Al investigar los tiempos que se requieren en el servicio en automóvil (en segundos), se obtuvieron los siguientes resultados (con base en datos del QSR Drive-Thru Time Study).

McDonald's:	287	128	92	267	176	240	192	118	153	254	193	136
Jack in the Box:	190	229	74	377	300	481	428	255	328	270	109	109

¿Cuál de los dos gigantes de comida rápida parece ser más veloz? ¿La diferencia parece ser significativa?

12. Anchura de cráneos Las anchuras máximas de muestras de cráneos egipcios de varones, que datan del 4000 a.C. y del 150 d.C. (de acuerdo con los datos de *Ancient Races of the Thebaid*, de Thomson y Randall-Maciver) se muestran a continuación:

4000 a.C.:	131	119	138	125	129	126	131	132	126	128	128	131
150 d.C.:	136	130	126	126	139	141	137	138	133	131	134	129

Los cambios del tamaño de las cabezas a través del tiempo sugieren la mezcla con personas de otras regiones. ¿Parece haber cambiado el tamaño de las cabezas del 4000 a.C. al 150 d.C.?

En los ejercicios 13 a 16, remítase a los conjuntos de datos del Apéndice B. Utilice un programa de cómputo o una calculadora para obtener las medias y las medianas, luego compare los resultados, tal como se indica.

13. Circunferencia de cabezas Para diagnosticar de forma correcta la hidrocefalia, un pediatra investiga la circunferencia de las cabezas de niños y niñas de dos años de edad. Utilice los resultados muestrales listados en el conjunto de datos 3. ¿Parece haber alguna diferencia entre los dos géneros?

14. Clancy, Rowling, Tolstoi Un psicólogo infantil investiga las diferencias en cuanto a la facilidad de la lectura; obtiene datos de *El oso y el dragón*, de Tom Clancy; *Harry Potter y la piedra filosofal*, de J. K. Rowling, y *La guerra y la paz*, de León Tolstoi. Remítase al conjunto de datos 14 del Apéndice B y utilice las puntuaciones de la calificación de Flesch-Kincaid de las 12 páginas seleccionadas aleatoriamente, para cada uno de los tres libros. ¿Los datos parecen ser diferentes?

15. Lluvia en el fin de semana Utilice el conjunto de datos 11 del Apéndice B para calcular la media y la mediana de las cantidades de lluvia que caen en Boston los jueves; calcule también, la media y la mediana de las cantidades de lluvia que caen en Boston los domingos. Los reportes de los medios de comunicación afirmaron que llueve más durante los fines de semana que entre semana. ¿Estos resultados apoyan dicha afirmación?

16. Consumo de tabaco/alcohol en películas infantiles En el artículo "Tobacco and Alcohol Use in G-Rated Children's Animated Films", de Goldstein, Sobel y Newman (*Journal of the American Medical Association*, vol. 281, núm. 12), se registraron las duraciones (en segundos) de escenas que muestran consumo de tabaco y alcohol en películas infantiles de dibujos animados. Remítase al conjunto de datos del Apéndice B, luego calcule la media y la mediana de las duraciones de escenas con tabaco así como la media y la mediana de las escenas con alcohol. ¿Parece haber una diferencia entre tales duraciones? ¿Cuál parece ser el problema mayor: las escenas que presentan consumo de tabaco o aquellas que muestran consumo de alcohol?

En los ejercicios 17 a 20 calcule la media de los datos que se resumen en la distribución de frecuencias dada.

17. Old Faithful Los visitantes del Parque Nacional Yellowstone consideran que una erupción del géiser Old Faithful es una gran atracción que uno no debe perderse. La distribución de frecuencias dada resume una muestra de los tiempos (en minutos) entre las erupciones.

Tabla del ejercicio 17

Tiempo	Frecuencia
40–49	8
50–59	44
60–69	23
70–79	6
80–89	107
90–99	11
100–109	1

18. Dado cargado El autor taladró un hoyo en un dado y lo relleno con plomo, después procedió a lanzarlo 200 veces. Los resultados se presentan en la distribución de frecuencias al margen. ¿El resultado parece ser muy diferente del resultado esperado con un dado inalterado?

Tabla del ejercicio 18

Resultado	Frecuencia
1	27
2	31
3	42
4	40
5	28
6	32

19. Infracciones de tránsito La distribución de frecuencias describe las velocidades de conductores infraccionados por la policía en la ciudad de Poughkeepsie. Los conductores viajaban por una zona con límite de velocidad de 30 millas/hora en Creek Road, que pasa por la universidad del autor. ¿Cómo se compara la media con el límite de velocidad de 30 mi/h?

Tabla del ejercicio 19

Velocidad	Frecuencia
42–45	25
46–49	14
50–53	7
54–57	3
58–61	1

20. Temperaturas corporales La distribución de frecuencias al margen resume una muestra de temperaturas corporales humanas. (Véanse las temperaturas de medianoche del segundo día, listadas en el conjunto de datos 4 del Apéndice B.) ¿Cómo se compara la media con el valor de 98.6°F, que es el valor que la mayoría de la gente supone como la media?

Tabla del ejercicio 20

Temperatura	Frecuencia
96.5–96.8	1
96.9–97.2	8
97.3–97.6	14
97.7–98.0	22
98.1–98.4	19
98.5–98.8	32
98.9–99.2	6
99.3–99.6	4

2-4 Más allá de lo básico

21. Media recortada Ya que la media es muy sensible a los valores extremos, decimos que no es una medida de tendencia central *resistente*. La **media recortada** es más resistente. Para calcular la media recortada del 10% de un conjunto de datos, primero se acomodan los datos en orden, después se elimina el 10% de los valores inferiores y el 10% de los valores superiores; finalmente, se calcula la media de los valores restantes. Para los pesos de los osos, en el conjunto de datos 9 del Apéndice B, calcule *a*) la media; *b*) la media recortada del 10%; *c*) la media recortada del 20%. ¿Cómo se comparan los resultados?

22. Media de medias Con el uso de un almanaque, un investigador obtiene la media del salario de maestros de cada estado de Estados Unidos. Suma los 50 valores y luego los divide entre 50 para obtener la media. ¿Será el resultado igual a la media nacional del salario de maestros? ¿Por qué?

23. Grados de libertad Diez valores tienen una media de 75.0. Nueve de los valores son 62, 78, 90, 87, 56, 92, 70, 70 y 93.

a. Calcule el décimo valor.

b. Necesitamos crear una lista de n valores que contenga una media específica conocida. Tenemos la libertad de seleccionar cualesquiera valores que deseemos para algunos de los n valores. ¿Cuántos de los n valores pueden asignarse libremente antes de determinar los valores restantes?

24. Datos censurados Se realizó un experimento para probar la vida de baterías de automóviles. El experimento se llevó a cabo durante un tiempo fijo de cinco años. (Se dice que la prueba se *censura* a los cinco años.) Los resultados muestrales (en años) son 2.5, 3.4, 1.2, 5+, 5+ (donde 5+ indica que la batería aún funcionaba al final del experimento). ¿Qué se concluye acerca de la vida media de las baterías?

25. Media ponderada Kelly Bell obtiene calificaciones parciales de 65, 83, 80 y 90. En el examen final recibe una calificación de 92. Calcule la media ponderada, si cada uno de los exámenes parciales cuenta el 15% y el examen final cuenta el 40% de la calificación total.

26. Datos transformados En cada uno de los siguientes casos, describa cómo se ven afectadas la media, la mediana, la moda y la mitad del rango.

a. La misma constante k se suma a cada valor del conjunto de datos.

b. Cada valor del conjunto de datos se multiplica por la misma constante k.

27. La **media armónica** se utiliza a menudo como una medida de tendencia central para conjuntos de datos que consisten en tasas de cambios, como la velocidad. Para calcularla, se divide el número de valores n entre la suma de los *recíprocos* de todos los valores, de la siguiente forma:

$$\frac{n}{\Sigma \frac{1}{x}}$$

(Ningún valor puede ser cero). Cuatro estudiantes conducen desde Nueva York hasta Florida (1,200 millas), a una velocidad de 40 mi/h (¡sí, como no!). Como necesitan llegar a su clase de estadística a tiempo, viajan de regreso a una velocidad de 60 mi/h. ¿Cuál es la velocidad promedio del viaje completo? (La media armónica se utiliza para promediar velocidades).

28. La **media geométrica** suele utilizarse en negocios y economía para calcular las tasas de cambio promedio, las tasas de crecimiento promedio o tasas promedio. Dados n valores (todos positivos), la media geométrica es la n-ésima raíz de su producto. El *factor de crecimiento promedio* de dinero compuesto con tasas de interés anual del 10%, el 8%, el 9%, el 12% y el 7% se obtiene determinando la media geométrica de 1.10, 1.08, 1.09, 1.12 y 1.07. Calcule el factor de crecimiento promedio.

29. La **media cuadrática** (o **cuadrado medio de raíz**, o **CMR**) suele utilizarse en aplicaciones físicas. Por ejemplo, en los sistemas de distribución de energía, los montajes y las corrientes suelen referirse en términos de sus valores de CMR. La media cuadrática de un conjunto de valores se obtiene elevando al cuadrado cada valor, sumando los resultados, dividiendo el número de valores n y después sacando la raíz cuadrada del resultado, el cual se expresa como

$$\text{media cuadrática} = \sqrt{\frac{\Sigma x^2}{n}}$$

Calcule el CMR de estas fuentes de poder (en volts): 110, 0, −60, 12.

30. Mediana Cuando los datos se resumen en una distribución de frecuencias, la mediana puede calcularse si primero se identifica la *clase de la mediana* (la clase que contiene a la mediana). Entonces, suponemos que los valores en esa clase se distribuyen uniformemente y podemos interpolar. Este proceso se describe de la siguiente forma:

$$(\text{límite inferior de clase de la mediana}) + (\text{anchura de clase})\left(\frac{\left(\frac{n+1}{2}\right) - (m+1)}{\text{frecuencia de clase de la mediana}}\right)$$

donde n es la suma de todas las frecuencias de clase y m es la suma de las frecuencias de clase que *preceden* la clase de la mediana. Utilice este procedimiento para calcular la mediana del conjunto de datos que se resume en la tabla 2.2. ¿Cómo se compara este resultado con la mediana de la lista original de datos, que es de 170? ¿Cuál valor de la mediana es mejor: el que se calculó para la tabla de frecuencias o el de 170?

2-5 Medidas de variación

Sugerencia: Ya que la sección introduce el concepto de variación, que es muy relevante en la estadística, es una de las más importantes de todo el libro. Primero lea esta sección en forma rápida y obtenga una comprensión general de las características de variación. Después, aprenda a calcular las medidas de variación, en especial la desviación estándar. Finalmente, trate de comprender el razonamiento que subyace a la fórmula de la desviación estándar, pero no gaste demasiado tiempo memorizando fórmulas o haciendo cálculos aritméticos. En lugar de ello, dé prioridad a aprender a *interpretar* los valores de la desviación estándar.

En la figura 2-12, de la página 74, se presenta un ejemplo visual de variación, el cual incluye muestras de tornillos de dos compañías diferentes. Puesto que dichos tornillos se utilizan para unir las alas al fuselaje, su calidad es muy importante. Si sólo tomamos en consideración la media, no reconoceríamos cualquier diferencia entre dos muestras, ya que ambas tienen una media de $\bar{x} = 2{,}000$ pulgadas. Sin embargo, debe ser evidente que las muestras difieren mucho con respecto a las *variaciones* de las longitudes de los tornillos. Los tornillos fabricados por Precision Bolt Co. parecen tener longitudes muy similares, mientras que las longitudes de los tornillos de Ruff Bolt Co. varían mucho. En muchos procesos de fabricación, este mismo aspecto tiene una gran importancia. Se logra una mejor calidad a través de una variación menor. En esta sección queremos que desarrolle la habilidad para medir y comprender la variación.

Otra situación ideal, que ilustra la importancia de la variación, se percibe en las filas de espera en los bancos. En el pasado, muchos bancos requerían que sus clientes esperaran en filas separadas, frente a cada una de las ventanillas; sin embargo, ahora la mayoría utiliza una sola fila de espera. ¿Por qué hicieron este cambio? El tiempo medio de espera no cambió, ya que la configuración de la fila de

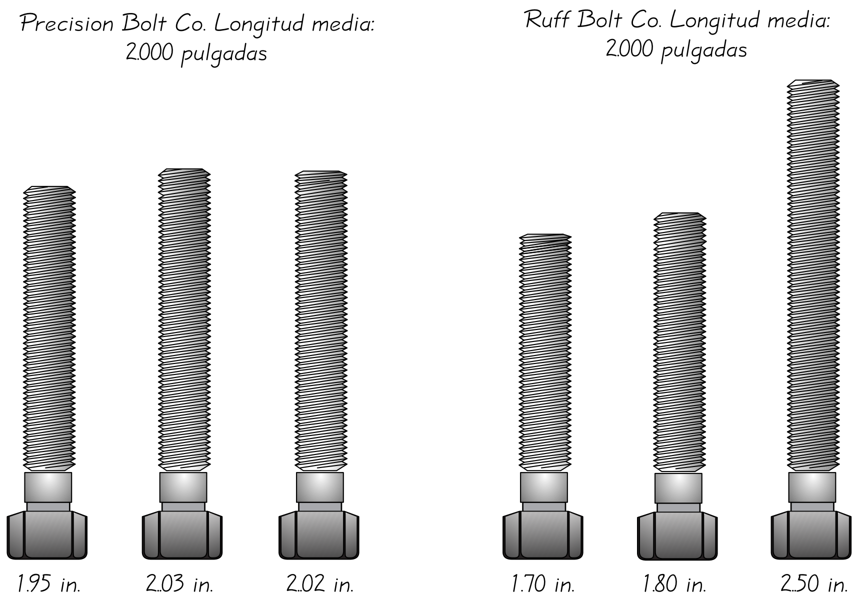


FIGURA 2-12 Tornillos fabricados por dos compañías diferentes

espera no afecta la eficiencia de los cajeros. El cambio a una sola fila se hizo porque los clientes prefieren esperar periodos que sean más *consistentes*, con menos variación. Miles de bancos hicieron un cambio que resultó en una menor variación (y clientes más contentos), aun cuando no se afectó la media. Consideremos algunos tiempos de espera específicos (en minutos) de clientes bancarios.

Banco Salem (una fila de espera)	4	7	7
Banco Mulberry (múltiples filas de espera)	1	3	14

Es fácil calcular que $\bar{x} = 6.0$ para ambos conjuntos de datos. También es fácil notar, mediante una inspección visual, que los tiempos de espera de 4, 7, 7 varían mucho menos que los tiempos de espera de 1, 3, 14. Procedamos ahora a desarrollar algunas formas específicas de medición real de la variación, de modo que sea posible utilizar números específicos en lugar de juicios subjetivos. Comencemos por el rango.

Rango

Definición

El **rango** de un conjunto de datos es la diferencia entre el valor máximo y el valor mínimo.

$$\textbf{Rango} = \textbf{(valor máximo)} - \textbf{(valor mínimo)}$$

Para calcular el rango, sólo se resta el valor mínimo del valor máximo. Para los clientes del Banco Salem, el rango es $7 - 4 = 3$ min. El Banco Mulberry tiene tiempos de espera con un rango de 13 min y este valor más grande sugiere una mayor variación.

Es muy fácil calcular el rango, pues depende únicamente de los valores máximo y mínimo, pero no es tan útil como otras medidas de variación que usan cada valor. (Véase el ejercicio 35 como un ejemplo en el que el rango causó confusión).

Desviación estándar de una muestra

La desviación estándar es, por lo general, la medida de variación más importante y útil. Definimos ahora la desviación estándar, pero para comprenderla por completo, es necesario estudiar el apartado “Interpretación y comprensión de la desviación estándar”, que aparece posteriormente en esta sección (véase la página 81).

Definición

Desviación estándar (de un conjunto de valores muestrales): medida de variación de los valores con respecto a la media. Es un tipo de desviación promedio de los valores, con respecto a la media, que se calcula utilizando las fórmulas 2-4 o 2-5.

Fórmula 2-4

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

desviación estándar de la muestra

Fórmula 2-5

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

fórmula abreviada de la desviación estándar de la muestra

Más adelante, en esta sección, analizaremos los fundamentos de estas fórmulas, aunque por ahora le recomendamos que utilice la fórmula 2-4, para resolver algunos ejemplos, y que después aprenda a calcular los valores de la desviación estándar por medio del uso de su calculadora y de un programa de cómputo. (La mayoría de las calculadoras científicas se diseñaron de tal modo que permiten que se introduzca una lista de valores y se obtenga de forma automática la desviación estándar). Mientras tanto, citamos propiedades importantes que son consecuencia de la forma en que se define la desviación estándar.

- La desviación estándar es una medida de variación de todos los valores con respecto a la *media*.
- El valor de la desviación estándar s suele ser positivo. Sólo es igual a cero cuando todos los valores de los datos son el mismo número. Además, valores grandes de s indican mayores cantidades de variación.
- El valor de la desviación estándar s se puede incrementar de manera drástica con la inclusión de uno o más datos distantes (valores de datos que se encuentran muy lejos de los demás).
- Las unidades de la desviación estándar s (como minutos, pies, libras, etcétera) son las mismas de los datos originales.

Más acciones, menos riesgo

En su libro *Investments*, Zvi Bodie, Alex Kane y Alan Marcus afirman que “la desviación estándar promedio de los rendimientos de carteras compuestas por un solo tipo de acciones fue de 0.554. El riesgo promedio de la cartera disminuye rápidamente cuando aumenta el número de acciones incluidas en la cartera”. También señalan que con 32 acciones la desviación estándar es de 0.325, lo que indica mucho menos variación y riesgo. Los autores destacan que, con sólo unas cuantas acciones, una cartera tiene alto grado de riesgo “específico de una empresa”, lo que significa que el riesgo puede atribuirse a la poca cantidad de acciones implicadas. Con más de 30 acciones, hay muy poco riesgo específico de una empresa; en su lugar, casi todo el riesgo es “riesgo de mercado”, atribuible al mercado de acciones global. Además, señalan que estos principios son “sólo una aplicación de la bien conocida ley de promedios”.

Procedimiento para calcular la desviación estándar con la fórmula 2-4

Paso 1: Calcule la media $\bar{x}$.

Paso 2: Reste la media de cada valor individual para tener una lista de desviaciones de la forma $(x - \bar{x})$.

Paso 3: Eleve al cuadrado cada una de las diferencias obtenidas en el paso 2. Esto produce números de la forma $(x - \bar{x})^2$.

Paso 4: Sume todos los cuadrados obtenidos en el paso 3. Éste es el valor de $\Sigma(x - \bar{x})^2$.

Paso 5: Divida el total del paso 4 entre el número $(n - 1)$, es decir, 1 menos que el total de valores presentes.

Paso 6: Calcule la raíz cuadrada del resultado del paso 5.

EJEMPLO Uso de la fórmula 2-4 Use la fórmula 2-4 para calcular la desviación estándar de los tiempos de espera de los clientes del Banco Mulberry. Dichos tiempos (en minutos) son 1, 3, 14.

SOLUCIÓN Utilizaremos los seis pasos en este proceso. Remítase a dichos pasos y a la tabla 2-11, que presenta los cálculos detallados.

Paso 1: Obtenga la media de 6.0 sumando los valores y después dividiendo entre el número de valores:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{18}{3} = 6.0 \text{ min}$$

Paso 2: Reste la media de 6.0 de cada valor para obtener los valores de $(x - \bar{x})$: $-5, -3, 8$.

Paso 3: Eleve al cuadrado cada valor que se obtuvo en el paso 2 para lograr valores de $(x - \bar{x})^2$: 25, 9, 64.

Paso 4: Sume todos los valores anteriores para obtener el valor de

$$\Sigma(x - \bar{x})^2 = 98$$

Paso 5: Con $n = 3$ valores, divida entre 1 menos que 3:

$$\frac{98}{2} = 49.0$$

Paso 6: Calcule la raíz cuadrada de 49.0. La desviación estándar es

$$\sqrt{49.0} = 7.0 \text{ min}$$

De manera ideal, ahora interpretaríamos el significado de los resultados; dichas interpretaciones se analizarán más tarde en esta sección.

Tabla 2-11 Cálculo de la desviación estándar de los tiempos de espera de los clientes del Banco Mulberry

x	$x - \bar{x}$	$(x - \bar{x})^2$
1	−5	25
3	−3	9
14	8	64
Totales: 18		98

$$\bar{x} = \frac{18}{3} = 6.0 \text{ min} \qquad s = \sqrt{\frac{98}{3-1}} = \sqrt{49} = 7.0 \text{ min}$$

EJEMPLO Uso de la fórmula 2-5 En el ejemplo anterior se utilizó la fórmula 2-4 para calcular la desviación estándar de los tiempos de espera de los clientes del Banco Mulberry. Con el mismo conjunto de datos, calcule la desviación estándar con la fórmula 2-5.

SOLUCIÓN La fórmula 2-5 requiere que primero encontremos valores para n, Σx y Σx^2.

$n = 3$ (ya que existen tres valores en la muestra)

$\Sigma x = 18$ (se obtiene al sumar los tres valores muestrales)

$\Sigma x^2 = 206$ (se obtiene al sumar los cuadrados de los valores muestrales, $1^2 + 3^2 + 14^2$)

Si usamos la fórmula 2-5, obtendremos

$$s = \sqrt{\frac{n(\Sigma x^2) - (\Sigma x)^2}{n(n-1)}} = \sqrt{\frac{3(206) - (18)^2}{3(3-1)}} = \sqrt{\frac{294}{6}} = 7.0 \text{ min}$$

Una actividad adecuada es detenerse aquí y calcular la desviación estándar de los tiempos de espera del Banco Salem. Siga los mismos procedimientos de los dos ejemplos anteriores y verifique que, para el Banco Salem, $s = 1.7$ min. (También será importante desarrollar la habilidad para obtener valores de desviaciones estándar con el uso de calculadoras y de programas de cómputo). Aun cuando las interpretaciones de tales desviaciones estándar se analizarán posteriormente, ahora las compararemos para darnos cuenta de que la desviación estándar de los tiempos de espera del Banco Salem (1.7 min) es mucho menor que la desviación estándar del Banco Mulberry (7.0 min). Esto apoya nuestra conclusión subjetiva de que los tiempos de espera del Banco Salem tienen una variación mucho menor que los tiempos del Banco Mulberry.

¿Dónde están los bateadores de 0.400?

El último beisbolista que bateó más de 0.400 fue Ted Williams, quien promedió 0.406 en 1941. Hubo promedios por arriba de 0.400 en 1876, 1879, 1887, 1894, 1895, 1896, 1897, 1899, 1901, 1911, 1920, 1922, 1924, 1925 y 1930, pero ninguno desde 1941. ¿Ya no existen grandes bateadores? Stephen Jay Gould, de la Universidad de Harvard, señaló que el promedio de bateo medio se ha mantenido estable en 0.260 durante aproximadamente 100 años, pero la desviación estándar disminuyó de 0.049 en la década de 1870 hasta 0.031 en la actualidad. Él argumenta que las estrellas de hoy son tan buenas como las del pasado, pero que los mejores lanzadores actuales mantienen promedios por debajo de 0.400.

Desviación estándar de una población

En nuestra definición de la desviación estándar, nos referimos a datos *muestrales*. Para calcular la desviación estándar σ (sigma minúscula) de una *población*, se utiliza una fórmula ligeramente diferente: en lugar de dividir entre $n - 1$, se hace entre el tamaño N de la población, como en la siguiente expresión:

$$\sigma = \sqrt{\frac{\Sigma(x - \mu)^2}{N}} \qquad \text{desviación estándar de la población}$$

Ya que generalmente usamos datos muestrales, a menudo utilizaremos la fórmula 2-4, en la cual dividimos entre $n - 1$. Muchas calculadoras dan tanto la desviación estándar muestral como la desviación estándar poblacional, pero con una gran variedad de notaciones diferentes. Asegúrese de identificar la notación de su calculadora, de modo que obtenga el resultado correcto. (La TI-83 Plus utiliza Sx para la desviación estándar muestral y σx para la desviación estándar poblacional).

Varianza de una muestra y una población

Usamos el término *variación* como una descripción general de la cantidad que varían los valores entre sí. (En ocasiones, se aplica el término *dispersión* en lugar de *variación*). El término *varianza* se refiere a una definición específica.

Definiciones

Varianza (de un conjunto de valores): medida de variación igual al cuadrado de la desviación estándar.

Varianza muestral: cuadrado de la desviación estándar s.

Varianza poblacional: cuadrado de la desviación estándar poblacional σ.

Se dice que la varianza muestral s^2 es un **estimador sin sesgo** de la varianza poblacional σ^2, lo que significa que los valores de s^2 tienden a igualar el valor de σ^2, en lugar de hacerlo de manera sistemática, a sobreestimar o subestimar σ^2. (Véase el ejercicio 41).

EJEMPLO Cálculo de la varianza En el ejemplo anterior, empleamos los tiempos de espera de los clientes del Banco Mulberry para descubrir que la desviación estándar está dada por $s = 7.0$ min. Calcule la varianza de esa misma muestra.

SOLUCIÓN Ya que la varianza es el cuadrado de la desviación estándar, obtenemos los resultados que se muestran abajo. Note que las unidades de los valores de los datos están dadas en minutos y que la desviación estándar es de 7.0 minutos; la varianza está dada en unidades de min^2.

$$\text{varianza muestral} = s^2 = 7.0^2 = 49.0 \text{ min}^2$$

La varianza es un estadístico importante que se utiliza en algunos métodos estadísticos relevantes, como el análisis de varianza, que se explica en el capítulo 11. Para nuestros propósitos presentes, la varianza tiene la siguiente gran desventaja:

las unidades de la varianza son diferentes a las unidades del conjunto original de datos. Por ejemplo, si los tiempos originales de espera de los clientes están dados en minutos, las unidades de varianza estarán dadas en minutos cuadrados ($\min^2$). ¿Qué es un minuto cuadrado? (Diviértase elaborando una respuesta creativa a dicha pregunta). Ya que la varianza utiliza unidades distintas, es sumamente difícil comprender la varianza si la se relaciona con el conjunto original de datos. Por esta propiedad, nos enfocaremos en la desviación estándar, mientras tratamos de comprender la variación.

Ahora presentamos la notación y la regla de redondeo que utilizamos.

Notación

s = desviación estándar *muestral*
s^2 = varianza *muestral*

σ = desviación estándar *poblacional*
σ^2 = varianza *poblacional*

Nota: Los artículos de las revistas y los reportes científicos suelen usar DE (o bien, SD en inglés) para la desviación estándar y VAR para la varianza.

Regla del redondeo

Usamos la misma regla de redondeo que se empleó en la sección 2-4:

> **Aumentar una posición decimal a la que había en los datos originales**.

Redondee sólo la respuesta final, no los valores a la mitad de un cálculo. (Si se vuelve absolutamente necesario redondear a la mitad, deberemos llevar al menos el doble de posiciones decimales de las que se utilizarán en la respuesta final).

Comparación de la variación en diferentes poblaciones

Anteriormente afirmamos que, ya que las unidades de la desviación estándar son las mismas que las unidades de los datos originales, es más fácil comprender la desviación estándar que la varianza. Sin embargo, esta misma propiedad dificulta comparar la variación de valores que se tomaron de distintas poblaciones. El *coeficiente de variación* resuelve tal desventaja.

Definición

Coeficiente de variación o **CV** de un conjunto de datos muestrales o poblacionales, expresado como porcentaje, describe la desviación estándar relativa a la media, y está dada de la siguiente forma:

Muestra	Población
$CV = \frac{s}{\bar{x}} \cdot 100\%$	$CV = \frac{\sigma}{\mu} \cdot 100\%$

EJEMPLO Estatura y peso de hombres Si utilizamos los datos muestrales estatura y peso de 40 hombres, incluidos en el conjunto de datos 1 del Apéndice B, obtendremos los estadísticos que aparecen en la siguiente tabla. Calcule el coeficiente de variación de las estaturas, después el coeficiente de variación de los pesos; finalmente, compare los dos resultados.

	Media ($\bar{x}$)	Desviación estándar (s)
Estatura	68.34 in	3.02 in
Peso	172.55 lb	26.33 lb

SOLUCIÓN Debido a que tenemos estadísticos muestrales, los dos coeficientes de variación se obtienen de la siguiente manera:

Estaturas: $$CV = \frac{s}{\bar{x}} \cdot 100\% = \frac{3.02 \text{ in}}{68.34 \text{ in}} \cdot 100\% = 4.42\%$$

Pesos: $$CV = \frac{s}{\bar{x}} \cdot 100\% = \frac{26.33 \text{ lb}}{172.55 \text{ lb}} \cdot 100\% = 15.26\%$$

Aun cuando la diferencia en unidades imposibilita la comparación de la desviación estándar de 3.02 pulgadas, con la desviación estándar de 26.33 libras, es posible comparar los coeficientes de variación, que carecen de unidades. Se observa que las estaturas (con $CV = 4.42\%$) tienen una variación considerablemente menor que los pesos (con $CV = 15.26\%$). Lo anterior tiene sentido, ya que, por lo general, vemos que los pesos de los hombres varían mucho más que sus estaturas. Por ejemplo, es muy raro encontrar un adulto que mida el doble que otro, pero es mucho más común ver a uno que pese el doble que otro.

Cálculo de la desviación estándar a partir de una distribución de frecuencias

En ocasiones necesitamos calcular la desviación estándar de un conjunto de datos que se resume en una distribución de frecuencias, como en la tabla 2-2 de la sección 2-2. Si se dispone de la lista original de valores muestrales, se utiliza la fórmula 2-4 o la 2-5, de modo que el resultado es más exacto. Si los datos originales no están disponibles, se utiliza uno de los dos métodos siguientes:

1. Si el número total de valores no es demasiado grande, trabaje con su calculadora o programa de cómputo e introduzca cada marca de clase tantas veces como el número de la frecuencia de clase.
2. Calcule la desviación estándar con la fórmula 2-6.

Fórmula 2-6

$$s = \sqrt{\frac{n\left[\Sigma(f \cdot x^2)\right] - \left[\Sigma(f \cdot x)\right]^2}{n(n-1)}}$$

desviación estándar para una distribución de frecuencias

Tabla 2-12 Cálculo de la desviación estándar a partir de una distribución de frecuencias

Cotinina	Frecuencia f	Marca de clase, x	$f \cdot x$	$f \cdot x^2$
0–99	11	49.5	544.5	26952.75
100–199	12	149.5	1794.0	268203.00
200–299	14	249.5	3493.0	871503.50
300–399	1	349.5	349.5	122150.25
400–499	2	449.5	899.0	404100.50
Totales:	$\Sigma f = 40$		$\Sigma(f \cdot x) = 7080$	$\Sigma(f \cdot x^2) = 1692910$

EJEMPLO Niveles de cotinina de fumadores Calcule la desviación estándar de los 40 valores que se resumen en la distribución de frecuencias de la tabla 2.2, considerando que no se dispone del conjunto original de datos.

SOLUCIÓN

Método 1: La tabla 2-12 tiene marcas de clase de 49.5, 149.5, 249.5, 349.5 y 449.5. Con una calculadora o un programa de cómputo introduzca el valor de 49.5, 11 veces (ya que la frecuencia de la primera clase es 11); introduzca 149.5, 12 veces y así sucesivamente. Obtenga la desviación estándar de este conjunto de 40 marcas de clase. El resultado debe ser 106.2.

Método 2: Utilice la fórmula 2-6. La aplicación de la fórmula 2-6 requiere que primero obtengamos los valores de n, $\Sigma(f \cdot x)$ y $\Sigma(f \cdot x^2)$. Después de obtener estos valores de la tabla 2-12, apliquemos la fórmula 2-6 de la siguiente manera:

$$s = \sqrt{\frac{n\left[\Sigma(f \cdot x^2)\right] - \left[\Sigma(f \cdot x)\right]^2}{n(n-1)}} = \sqrt{\frac{40[1692910] - [7080]^2}{40(40-1)}}$$

$$= \sqrt{\frac{17{,}590{,}000}{1560}} = \sqrt{11275.64103} = 106.2$$

Calculadora TI-83 Plus A diferencia de la mayoría de las calculadoras, la TI-83 Plus calcula la desviación estándar de valores resumidos en una distribución de frecuencias. Primero, introduzca las marcas de clase en la lista L1, después introduzca las frecuencias en la lista L2. Ahora presione **STAT**, seleccione **CALC**, luego seleccione **1-VarStats** e introduzca L1, L2 para obtener resultados que incluyan la media y la desviación estándar. Nuevamente, la desviación estándar muestral se identifica con Sx y la desviación estándar poblacional con σx.

Interpretación y comprensión de la desviación estándar

Este apartado es sumamente importante, puesto que ahora trataremos de que la desviación estándar tenga sentido. Primero, debemos comprender con claridad que la desviación estándar mide la variación entre los valores. Los valores cercanos producirán una desviación estándar pequeña, mientras que los valores muy dispersos producirán una desviación estándar más grande.

Ya que la variación es un concepto tan importante y que la desviación estándar es una herramienta tan útil para medir la variación, consideraremos tres formas diferentes para lograr una apreciación de los valores de las desviaciones estándar. La primera es la **regla práctica del intervalo**, que se basa en el principio de que para muchos conjuntos de datos, la vasta mayoría (tanto como el 95%) de los valores muestrales se ubican dentro de dos desviaciones estándar de la media. (Es posible mejorar la precisión de tal regla si tomamos en cuenta factores como el tamaño de la muestra y la naturaleza de la distribución, aunque preferimos sacrificar precisión en aras de la sencillez. Además, podríamos usar tres o, incluso, cuatro desviaciones estándar en lugar de 2, lo cual constituye una decisión un poco arbitraria. Sin embargo, deseamos una regla sencilla que nos ayude a interpretar los valores de las desviaciones estándar; métodos posteriores producirán resultados más precisos).

Regla práctica del intervalo

Para estimar el valor de la desviación estándar s: para obtener un estimado burdo de la desviación estándar, utilice

$$s \approx \frac{\text{rango}}{4}$$

donde el rango = (valor máximo) − (valor mínimo).

Para interpretar un valor conocido de la desviación estándar: si se conoce la desviación estándar s, utilícela para calcular estimados burdos de los valores muestrales mínimos y máximos “comunes” por medio de

$$\text{valor mínimo “común”} \approx (\text{media}) - 2 \times (\text{desviación estándar})$$

$$\text{valor máximo “común”} \approx (\text{media}) + 2 \times (\text{desviación estándar})$$

Cuando calcule una desviación estándar por medio de las fórmulas 2-4 o 2-5, la regla práctica del intervalo resulta útil para verificar el resultado, pero debe estar consciente de que, aun cuando la aproximación nos acerca a la respuesta, puede tener un error considerable.

EJEMPLO Niveles de cotinina de fumadores Utilice la regla práctica del intervalo para calcular un estimado burdo de la desviación estándar de la muestra de 40 niveles de cotinina de fumadores, como se observa en la tabla 2-1.

SOLUCIÓN Al emplear la regla práctica del intervalo para estimar la desviación estándar de datos muestrales, calculamos el rango y lo dividimos entre 4. Si observamos la lista de los niveles de cotinina, notaremos que el mínimo es 0 y el máximo 491; por lo tanto, el rango es de 491. La desviación estándar s se estima de la siguiente manera:

$$s \approx \frac{\text{rango}}{4} = \frac{491}{4} = 122.75 \approx 123$$

INTERPRETACIÓN Este resultado es muy cercano al valor correcto de 119.5, que se obtiene al calcular el valor exacto de la desviación estándar con las fórmulas 2-4 o 2-5. No espere que la regla práctica del intervalo funcione tan bien en otros casos.

El siguiente ejemplo es particularmente importante como ilustración de una forma de *interpretar* el valor de una desviación estándar.

EJEMPLO **Circunferencias de la cabeza de niñas** Resultados anteriores del National Health Survey sugieren que las circunferencias de las cabezas de niñas de dos meses de edad tienen una media de 40.05 cm y una desviación estándar de 1.64 cm. Utilice la regla práctica del intervalo para calcular el mínimo y el máximo "comunes" de las circunferencias de las cabezas. (Estos resultados serían prácticos para un médico al que le interese identificar circunferencias "infrecuentes", que serían el resultado de un trastorno como la hidrocefalia). Después, determine si una circunferencia de 42.6 cm sería considerada "infrecuente".

SOLUCIÓN Con una media de 40.05 cm y una desviación estándar de 1.64 cm, empleamos la regla práctica del intervalo para calcular las circunferencias mínima y máxima comunes, de la siguiente manera:

$$\begin{aligned}\text{mínimo} &\approx (\text{media}) - 2 \times (\text{desviación estándar})\\ &= 40.05 - 2(1.64) = 36.77 \text{ cm}\end{aligned}$$

$$\begin{aligned}\text{máximo} &\approx (\text{media}) + 2 \times (\text{desviación estándar})\\ &= 40.05 + 2(1.64) = 43.33 \text{ cm}\end{aligned}$$

INTERPRETACIÓN Con base en estos resultados, esperamos que, generalmente, las niñas de dos meses de edad tengan una cabeza cuya circunferencia mida entre 36.77 cm y 43.33 cm. Como 42.6 cm está dentro de estos límites, se consideraría una niña normal.

Regla empírica para datos con distribución normal (o 68-95-99.7)

Otra regla útil para interpretar los valores de una desviación estándar es la **regla empírica**. Esta regla establece que las siguientes propiedades se aplican *a conjuntos de datos que tienen una distribución aproximadamente normal*. (Véase la figura 2-13).

- Aproximadamente el 68% de todos los valores están dentro de una desviación estándar de la media.
- Aproximadamente el 95% de todos los valores están dentro de dos desviaciones estándar de la media.
- Aproximadamente el 99.7% de todos los valores están dentro de tres desviaciones estándar de la media.

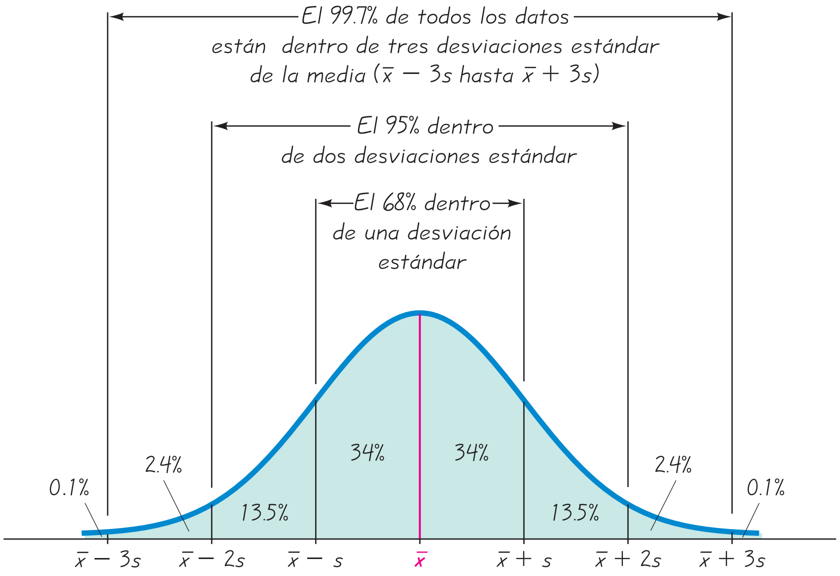


FIGURA 2-13 La regla empírica

EJEMPLO Puntuaciones de CI Las puntuaciones de adultos normales en la prueba Weschler tienen una distribución normal, con una media de 100 y una desviación estándar de 15. ¿Qué porcentaje de adultos tienen puntuaciones de CI entre 55 y 145?

SOLUCIÓN La clave para resolver el problema es reconocer que 55 y 145 están exactamente a tres desviaciones estándar de la media de 100, como se indica abajo.

$$3 \text{ desviaciones estándar} = 3s = 3(15) = 45$$

Por lo tanto, tres desviaciones estándar de la media son igual a

$$100 - 45 = 55$$

o

$$100 + 45 = 145$$

La regla empírica nos indica que aproximadamente el 99.7% de todos los valores están dentro de tres desviaciones estándar de la media; por lo tanto, el 99.7% de todas las puntuaciones de CI se encuentran entre 55 y 145.

Sugerencia: Las dificultades para aplicar la regla empírica suelen surgir de la confusión al interpretar frases tales como “dentro de 3 desviaciones estándar de la media”. Deténgase aquí, revise el ejemplo anterior hasta que el significado de dicha frase esté claro. Además, observe las siguientes interpretaciones generales de esa frase.

Frase	Significado
Dentro de una desviación estándar de la media	Entre $(\bar{x} - s)$ y $(\bar{x} + s)$
Dentro de dos desviaciones estándar de la media	Entre $(\bar{x} - 2s)$ y $(\bar{x} + 2s)$
Dentro de tres desviaciones estándar de la media	Entre $(\bar{x} - 3s)$ y $(\bar{x} + 3s)$

Un tercer concepto útil para comprender el valor de una desviación estándar es el **teorema de Chebyshev**. La regla empírica anterior se aplica sólo a conjuntos de datos con una distribución normal. El teorema de Chebyshev, en lugar de limitarse a conjuntos de datos con distribuciones normales se aplica a cualquier conjunto de datos, pero sus resultados son muy aproximados.

Teorema de Chebyshev

La proporción (o fracción) de cualquier conjunto de datos que está dentro de K desviaciones estándar de la media es siempre *al menos* $1 - 1/K^2$, donde K es cualquier número positivo mayor que 1. Para $K = 2$ y $K = 3$, tenemos los siguientes enunciados:

- Al menos $3/4$ (o 75%) de todos los valores están dentro de dos desviaciones estándar de la media.
- Al menos $8/9$ (u 89%) de todos los valores están dentro de tres desviaciones estándar de la media.

EJEMPLO Puntuaciones de CI Las puntuaciones de CI de adultos normales tomadas de la prueba Weschler tienen una media de 100 y una desviación estándar de 15. ¿Qué podemos concluir a partir del teorema de Chebyshev?

SOLUCIÓN Al aplicar el teorema de Chebyshev, con una media de 100 y una desviación estándar de 15, llegamos a las siguientes conclusiones:

- Por lo menos $3/4$ (o 75%) de todos los adultos tienen puntuaciones de CI que están dentro de dos desviaciones estándar de la media (entre 70 y 130).
- Al menos $8/9$ (u 89%) de todos los adultos tienen puntuaciones de CI que están dentro de tres desviaciones estándar de la media (entre 55 y 145).

Cuando intentemos darle un significado a un valor de una desviación estándar, debemos usar uno o más de los tres conceptos anteriores. Para comprender aún mejor la naturaleza de la desviación estándar, consideraremos los fundamentos subyacentes que conducen a la fórmula 2-4, que es la base de su definición. (La fórmula 2-5 es sencillamente otra versión de la fórmula 2-4, derivada de modo que los cálculos aritméticos pueden simplificarse).

Fundamentos de la fórmula 2-4

La desviación estándar de un conjunto de datos muestrales se define con las fórmulas 2-4 y 2-5, las cuales son equivalentes en el sentido de que siempre producen el mismo resultado. La fórmula 2-4 tiene la ventaja de reforzar el concepto de que la desviación estándar es un tipo de desviación promedio. La fórmula 2-5, la de ser más fácil de usar cuando hay que calcular desviaciones estándar por nuestra cuenta. La fórmula 2-5 también elimina los errores de redondeo intermedios que se introducen en la fórmula 2-4 cuando no se utiliza el valor exacto de la media. La fórmula 2-5 se aplica en calculadoras y programas, ya que requiere sólo de tres lugares de memoria (para n, $\sum x$ y $\sum x^2$), en vez de un lugar de memoria para cada valor del conjunto de datos.

¿Para qué definir una medida de variación en la forma descrita por la fórmula 2.4? Al medir la variación en un conjunto de datos muestrales, parece lógico iniciar con las cantidades individuales con las que los valores se desvían de la media. Para un valor particular x, la cantidad de **desviación** es $x - \bar{x}$, que es la diferencia entre el valor individual x y la media. Para los tiempos de espera del Banco Mulberry de 1, 3, 14, la media es 6.0, de modo que las desviaciones de la media son -5, -3 y 8. Sería bueno combinar, de alguna forma, dichas desviaciones en un solo valor colectivo. La simple suma de las desviaciones no funciona, ya que la suma siempre será cero. Para obtener un estadístico que mida la variación, necesitamos evitar la cancelación de números positivos y negativos. Un método consiste en sumar valores absolutos, como en $\Sigma|x - \bar{x}|$. Si calculamos la media de esta suma, obtendremos la **desviación media absoluta** (o **DMA**), que es la distancia media de los datos con respecto a la media.

$$\text{desviación media absoluta} = \frac{\Sigma|x - \bar{x}|}{n}$$

Ya que los tiempos de espera del Banco Mulberry de 1, 3, 14 tienen desviaciones de -5, -3 y 8, la desviación media absoluta es $(5 + 3 + 8)/3 = 6/3 = 5.3$.

¿Por qué no utilizar la desviación media absoluta? Como la desviación media absoluta requiere que usemos valores absolutos, emplea una operación que no es algebraica. (Las operaciones algebraicas incluyen la suma, la multiplicación, la raíz cuadrada y la elevación a potencias enteras o fraccionarias, pero el valor absoluto no está incluido). El uso de valores absolutos crea problemas algebraicos en los métodos inferenciales de la estadística. Por ejemplo, la sección 8.3 presentó un método para hacer inferencias acerca de las medias de dos poblaciones; dicho método se construye alrededor de una propiedad de adición de las varianzas, pero la desviación media absoluta no posee tal propiedad de adición. (He aquí una versión simplificada de la propiedad de adición de la varianza: si se tienen dos poblaciones independientes y se selecciona aleatoriamente un valor de cada población y se suman, dichas sumas tendrán una varianza que es igual a la suma de las varianzas de las dos poblaciones.) La misma propiedad de adición subyace en los fundamentos de la regresión, que se presentan en el capítulo 9, y el análisis de varianza que se introduce en el capítulo 11. Además, el ejercicio 42 demuestra que el valor de la media absoluta presenta un sesgo, lo cual significa que cuando se calculan valores de media absoluta de muestras, no se tiende a igualar el valor medio absoluto de la población. En contraste, la desviación estándar utiliza sólo operaciones algebraicas. Puesto que se basa en la raíz cuadrada de una suma de cuadrados, la desviación estándar se asemeja a las fórmulas de distancia que se usan en álgebra. Hay muchos ejemplos en los cuales un procedimiento estadístico se encuentra sesgado en una suma de cuadrados similar. Por lo tanto, en lugar de emplear valores absolutos, obtenemos una mejor medida de variación si logramos que todas las desviaciones $(x - \bar{x})$ no sean negativas, lo cual haremos elevándolas al cuadrado; este método conduce a la desviación estándar. Por tales razones, las calculadoras científicas suelen incluir una función para la desviación estándar, pero casi nunca la desviación media absoluta.

¿Por qué dividir entre $n - 1$? Después de obtener todos los valores individuales de $(x - \bar{x})^2$, los combinamos calculando su suma y luego obtenemos un promedio dividiéndola entre $n - 1$. Dividimos entre $n - 1$, porque hay solamente $n - 1$ valores independientes. Es decir, con una media dada, sólo a $n - 1$ valores se le puede asignar un número con libertad, antes de que se determine el último valor. Véase el ejercicio 41, que proporciona números concretos que ilustran como tal división entre $n-1$ es mejor que la división entre n. Este ejercicio muestra que si s^2 se definiera con la división entre n, de forma sistemática subestimaría el valor de σ^2, por lo que lo compensamos incrementando su valor general, haciendo que su denominador sea más pequeño (usando $n-1$ en lugar de n). El ejercicio 41 demuestra cómo la división entre $n-1$ provoca que la varianza muestral s^2 iguale el valor de la varianza poblacional σ^2; en tanto que la división entre n causa que la varianza muestral s^2 subestime el valor de la varianza poblacional σ^2.

El paso 6 de la fórmula 2-4, para el cálculo de una desviación estándar, implica sacar una raíz cuadrada. Esto se hace para compensar la elevación al cuadrado que se realizó en el paso 3. Una consecuencia importante de la obtención de la raíz cuadrada es que la desviación estándar tiene las mismas unidades de medición que los valores originales. Por ejemplo, si el tiempo de espera de los clientes se da en minutos, la desviación estándar de dichos tiempos también estará en minutos. Si nos detuviéramos en el paso 5, el resultado estaría dado en unidades de "minutos cuadrados", que es un concepto abstracto sin relación directa con la realidad.

Después de estudiar dicha sección, usted debe comprender que la desviación estándar es una medida de variación entre valores. Al tener datos muestrales, será capaz de calcular el valor de la desviación estándar y de interpretar los valores de las desviaciones estándar que obtuvo. Debe saber que, para un conjunto de datos común, es raro que un valor difiera de la media por más de 2 o 3 desviaciones estándar.

2-5 Destrezas y conceptos básicos

En los ejercicios 1 a 8 calcule el rango, la varianza y de desviación estándar para los datos muestrales dados. (En la sección 2-4 se utilizaron los mismos datos para calcular medidas de tendencia central. Aquí calculamos medidas de variación).

1. Consumo de tabaco en películas infantiles En el artículo "Tobbaco and Alcohol Use in G-Rated Children's Animated Films", de Goldstein, Sobel y Newman (*Journal of the American Medical Association*, vol. 281, núm. 12), se registró la duración (en segundos) de escenas de películas de dibujos animados, de los Universal Studios, que muestran consumo de tabaco. A continuación se presentan los primeros seis valores, que se incluyen en el conjunto de datos 7 del Apéndice B. ¿Parecen tales duraciones ser consistentes o varían ampliamente?

0	223	0	176	0	548

2. Harry Potter En un intento por medir el nivel de lectura de un libro, se obtuvieron los puntajes de la facilidad de lectura de Flesch de 12 páginas que se seleccionaron aleatoriamente, de la obra *Harry Potter y la piedra filosofal*, de J. K. Rowling. Dichos puntajes, que se encuentran en el conjunto de datos 14 del Apéndice B, se listan a continuación. Debido a que tales puntajes se basan en 12 páginas que se seleccionaron aleatoriamente, ¿es probable que la desviación estándar de esta muestra sea un estimado razonable de la desviación estándar de los niveles de lectura de todas las páginas del libro?

85.3	84.3	79.5	82.5	80.2	84.6
79.2	70.9	78.6	86.2	74.0	83.7

3. Cereal Un nutriólogo obtiene las cantidades de azúcar (en gramos) de un gramo de 16 cereales diferentes, incluyendo Cheerios, Corn Flakes, Fruit Loops, Trix y 12 más. Estos valores, que se incluyen en el conjunto de datos 16 del Apéndice B, se listan a continuación. ¿Será la desviación estándar de dichos valores un buen estimado de la desviación estándar de la cantidad de azúcar en cada gramo del cereal consumido por la población de todos los estadounidenses? ¿Por qué?

0.03	0.24	0.30	0.47	0.43	0.07	0.47	0.13
0.44	0.39	0.48	0.17	0.13	0.09	0.45	0.43

4. Índice de masa corporal Como parte del National Health Examination se mide el índice de masa corporal en una muestra aleatoria de mujeres. Algunos de los valores, que se anexan en el conjunto de datos 1 del Apéndice B, se listan a continuación. ¿Estará la desviación estándar de la muestra razonablemente cerca de la desviación estándar de 6.17, que es la desviación estándar de las 40 mujeres que se incluyen en el conjunto de datos 1?

19.6	23.8	19.6	29.1	25.2	21.4	22.0	27.5
33.5	20.6	29.9	17.7	24.0	28.9	37.7	

5. Conductores alcoholizados Abajo se listan las concentraciones de alcohol en la sangre de conductores que se vieron envueltos en accidentes fatales y que después fueron sentenciados a prisión (de acuerdo con datos del Departamento de Justicia de Estados Unidos). Cuando un estado lanza una campaña para "reducir el número de conductores alcoholizados", ¿es la intención de la campaña disminuir la desviación estándar?

0.27	0.17	0.17	0.16	0.13	0.24	0.29	0.24
0.14	0.16	0.12	0.16	0.21	0.17	0.18	

6. Muertes en motocicleta A continuación se presentan las edades de motociclistas cuando se accidentaron fatalmente en accidentes de tránsito (de acuerdo con datos del Departamento del Transporte de Estados Unidos). ¿De qué manera se compara la variación de estas edades con la variación de las edades de conductores con licencia en la población general?

17	38	27	14	18	34	16	42	28
24	40	20	23	31	37	21	30	25

7. Tiempos de reacción El autor visitó el Museo de Ciencias Reuben H. Fleet, en San Diego, y repitió un experimento de tiempos de reacción. Se obtuvieron los siguientes tiempos (en centésimas de segundo). ¿De que manera las medidas de variación reflejan el hecho de que tales tiempos parezcan muy consistentes?

19	20	17	21	21	21	19	18	19	19
17	17	15	17	18	17	18	18	18	17

8. Tabletas de Bufferin A continuación se listan los pesos medidos (en miligramos) de una muestra de tabletas de aspirina Bufferin. Como este medicamento debe fabricarse de forma consistente para que las dosis se controlen, ¿las medidas de variación parecen indicar que la variación tiene un nivel aceptable?

672.2	679.2	669.8	672.6	672.2	662.2
662.7	661.3	654.2	667.4	667.0	670.7

En los ejercicios 9 a 12 calcule el rango, la varianza y la desviación estándar para cada una de las dos muestras; luego, compare los dos conjuntos de resultados. (En la sección 2.4 se utilizaron los mismos datos).

9. Tiempos de espera de clientes A continuación se presentan los tiempos de espera (en minutos) de los clientes del Banco Jefferson Valley (donde todos los clientes forman una sola fila) y del Banco Providence (donde los clientes esperan en filas individuales, en tres ventanillas diferentes):

Jefferson Valley:	6.5	6.6	6.7	6.8	7.1	7.3	7.4	7.7	7.7	7.7
Providence:	4.2	5.4	5.8	6.2	6.7	7.7	7.7	8.5	9.3	10.0

10. Coca Cola regular/Coca Cola dietética Los siguientes son los pesos (en libras) de muestras del contenido de latas de Coca Cola regular y Coca Cola dietética:

Regular:	0.8192	0.8150	0.8163	0.8211	0.8181	0.8247
Dietética:	0.7773	0.7758	0.7896	0.7868	0.7844	0.7861

11. Mickey D *vs.* Jack Al investigar los tiempos que se requieren en el servicio para automóvil (en segundos), se obtienen los siguientes resultados (con base en datos del QSR Drive-Thru Time Study).

McDonald's:	287	128	92	267	176	240	192	118	153	254	193	136
Jack in the Box:	190	229	74	377	300	481	428	255	328	270	109	109

12. Anchura de cráneos Las anchuras máximas de muestras de cráneos egipcios de varones que datan del 4000 a.C. y del 150 d.C (de acuerdo con datos de *Ancient Races of the Thebaid*, de Thomson y Randall-Maciver) se muestran a continuación:

4000 a.C.:	131	119	138	125	129	126	131	132	126	128	128	131
150 d.C.:	136	130	126	126	139	141	137	138	133	131	134	129

En los ejercicios 13 a 16 remítase a los conjuntos de datos del Apéndice B. Utilice un programa de cómputo o una calculadora para obtener las desviaciones estándar; luego, compare los resultados.

T **13.** Circunferencia de cabezas Para diagnosticar de forma correcta el trastorno de hidrocefalia, un pediatra investiga la circunferencia de las cabezas de niños y niñas de dos años de edad. Utilice los resultados muestrales listados en el conjunto de datos 3. ¿Hay alguna diferencia entre los dos géneros?

T **14.** Clancy, Rowling, Tolstoi Un psicólogo infantil investiga las diferencias en la facilidad de lectura; obtiene datos con *El oso y el dragón*, de Tom Clancy; *Harry Potter y la piedra filosofal*, de J. K. Rowling, y *La guerra y la paz*, de León Tolstoi. Remítase al conjunto de datos 14 del Apéndice B y utilice las puntuaciones de la calificación de Flesch-Kincaid de las 12 páginas seleccionadas aleatoriamente para cada uno de los tres libros.

T **15.** Lluvia el fin de semana Utilice el conjunto de datos 11 del Apéndice B, sobre las cantidades de lluvia que caen en Boston los jueves y las que caen los domingos.

T **16.** Consumo de tabaco/alcohol en películas infantiles En el artículo "Tobacco and Alcohol Use in G-Rated Children's Animated Films", de Goldstein, Sobel y Newman (*Journal of the American Medical Association*, vol. 281, núm. 12), se registraron las duraciones (en segundos) de escenas que muestran consumo de tabaco y alcohol en películas infantiles de dibujos animados. En el conjunto de datos 7 del Apéndice B, utilice las duraciones de escenas con tabaco y después las escenas con alcohol.

En los ejercicios 17 a 20 calcule la desviación estándar de los datos que se resumen en la distribución de frecuencias dada. (En la sección 2.4 se utilizaron las mismas distribuciones de frecuencias.)

17. Old Faithful Los visitantes del Parque Nacional Yellowstone consideran que una erupción del géiser Old Faithful es una gran atracción que uno no debe perderse. La distribución de frecuencias dada resume una muestra de los tiempos (en minutos) entre las erupciones.

18. Dado cargado El autor taladró un hoyo en un dado y lo rellenó con plomo; después, procedió a lanzarlo 200 veces. Los resultados se presentan en la distribución de frecuencias al margen.

19. Infracciones de tránsito La distribución de frecuencias describe las velocidades de conductores infraccionados por la policía en la ciudad de Poughkeepsie. Los conductores viajaban a través de una zona con límite de velocidad de 30 millas/hora en Creek Road, que pasa por la universidad del autor.

Tabla del ejercicio 17

Tiempo	Frecuencia
40–49	8
50–59	44
60–69	23
70–79	6
80–89	107
90–99	11
100–109	1

Tabla del ejercicio 18

Resultado	Frecuencia
1	27
2	31
3	42
4	40
5	28
6	32

Tabla del ejercicio 19

Velocidad	Frecuencia
42–45	25
46–49	14
50–53	7
54–57	3
58–61	1

Tabla del ejercicio 20

Temperatura	Frecuencia
96.5–96.8	1
96.9–97.2	8
97.3–97.6	14
97.7–98.0	22
98.1–98.4	19
98.5–98.8	32
98.9–99.2	6
99.3–99.6	4

20. Temperaturas corporales La distribución de frecuencias al margen resume una muestra de temperaturas corporales humanas. (Véanse las temperaturas de medianoche del segundo día, listadas en el conjunto de datos 4 del Apéndice B).

21. Edades de profesores Utilice la regla práctica del intervalo para estimar la desviación estándar de las edades de todos los profesores de su universidad.

22. Calificaciones de prueba Con la regla práctica del intervalo estime la desviación estándar de las calificaciones del primer examen de estadística de su clase.

23. Longitudes de piernas En los datos muestrales del conjunto de datos 1 del Apéndice B, las longitudes del muslo de la muestra de 40 mujeres tienen una media de 38.86 centímetros y una desviación estándar de 3.78 centímetros. Use la regla práctica del intervalo para estimar las longitudes mínima y máxima "comunes" de los muslos de las mujeres. En dicho contexto, ¿una longitud de 47.0 centímetros sería considerada infrecuente?

24. Estaturas de mujeres La media de las estaturas de las mujeres es de 63.6 pulgadas, con una desviación estándar de 2.5 pulgadas (con base en datos del National Health Survey). Utilice la regla práctica del intervalo para estimar las estaturas mínima y máxima "comunes" de las mujeres. En tal contexto, ¿es poco común que una mujer mida seis pies?

25. Estaturas de mujeres Las estaturas de las mujeres tienen una distribución normal, con una media de 63.6 pulgadas y una desviación estándar de 2.5 pulgadas. Utilice la regla empírica para determinar el porcentaje aproximado de mujeres que están entre
a. 61.1 y 66.1 pulgadas
b. 56.1 y 71.1 pulgadas

26. Pesos de la Coca Cola regular Con los pesos de la Coca Cola regular listados en el conjunto de datos 17 del Apéndice B, encontramos que la media es de 0.81682 libras, la desviación estándar es de 0.00751 libras y la distribución es aproximadamente normal. Aplique la regla empírica y determine el porcentaje aproximado de latas de Coca Cola regular que tienen pesos entre
a. 0.80931 y 0.82433 libras
b. 0.80180 y 0.83184 libras

27. Estaturas de mujeres Si las estaturas de mujeres tienen una media de 63.6 pulgadas y una desviación estándar de 2.5 pulgadas, ¿qué se concluye a partir del teorema de Chebyshev acerca del porcentaje de mujeres que están entre 58.6 pulgadas y 68.6 pulgadas?

28. Pesos de la Coca Cola regular Utilizando los pesos de la Coca Cola regular listados en el conjunto de datos 17 del Apéndice B, encontramos que la media es de 0.81682 libras y la desviación estándar es de 0.00751 libras. ¿Qué concluye a partir del teorema de Chebyshev acerca del porcentaje de latas de Coca Cola regular con pesos que están entre 0.79429 y 0.83935 libras?

T **29.** Coeficiente de variación del cereal Remítase al conjunto de datos 17 del Apéndice B. Calcule el coeficiente de variación de las calorías y el coeficiente de variación de los gramos de azúcar por gramo de cereal. Compare los resultados.

T **30.** Coeficiente de variación de Coca Cola y de Pepsi Remítase al conjunto de datos 17 del Apéndice B. Calcule el coeficiente de variación de los pesos de la Coca Cola regular y después el coeficiente de variación de los pesos de la Pepsi regular. Compare los resultados. ¿Alguna de las dos compañías parece tener pesos significativamente más consistentes?

31. Igualdad para todos ¿Qué sabe usted acerca de los valores en un conjunto de datos con una desviación estándar $s = 0$?

32. Comprensión de las unidades de medición Si un conjunto de datos consiste en multas por exceso de velocidad (en dólares), ¿qué unidades se utilizan para la desviación estándar? ¿Qué unidades se utilizan para la varianza?

33. Comparación de baterías para automóviles Las marcas de baterías para automóviles Everlast y Endurance aseguran en su etiqueta una duración de 48 meses. En realidad, ambas tienen una vida media de 50 meses, pero las baterías Everlast tienen una desviación estándar de dos meses, mientras que la de las baterías Endurance es de seis meses. ¿Cuál de las marcas sería una mejor opción? ¿Por qué?

34. Interpretación de datos distantes Un conjunto de datos consta de 20 valores, bastante cercanos entre sí. Se incluye otro valor, pero éste es un dato distante (muy lejos de los demás). ¿De qué manera se ve afectada la desviación estándar por el dato distante? ¿No genera efecto alguno? ¿Tiene un efecto pequeño? ¿Un efecto grande?

2-5 Más allá de lo básico

35. Comparación de conjuntos de datos Dos secciones diferentes de un curso de estadística resuelven el mismo examen, cuyas calificaciones se muestran abajo. Calcule el rango y la desviación estándar de cada sección. ¿Qué se concluye acerca de la variación en las dos secciones, a partir de los valores del rango? ¿Por qué el rango causa confusión en este caso? ¿Qué se concluye acerca de la variación en las dos secciones con respecto a los valores de la desviación estándar?

Sección 1:	1	20	20	20	20	20	20	20	20	20	20
Sección 2:	2	3	4	5	6	14	15	16	17	18	19

36. Transformación de datos Describa de qué forma se afectan el rango y la desviación estándar de un conjunto de datos en los siguientes casos:

a. Se suma la misma constante K a cada valor del conjunto de datos.

b. Cada valor del conjunto de datos se multiplica por la misma constante K.

c. Para los datos de temperaturas corporales listados en el conjunto de datos 4 del Apéndice B (12 a. m. del día 2), $\bar{x} = 98.20°\text{F}$ y $s = 0.62°\text{F}$. Calcule los valores de y s después de convertir cada temperatura a grados Celsius. [*Considere que* $C = 5\,(F - 32)/9$].

37. Genichi Taguchi creó un método para mejorar la calidad y reducir los costos de fabricación, por medio de una combinación de ingeniería y estadística. Una herramienta básica en el método de Taguchi es el **cociente señal-ruido**. La forma más simple para calcular tal cociente es dividir la media entre la desviación estándar. Calcule el cociente señal-ruido de los niveles de cotinina de fumadores listados en la tabla 2-1.

38. Sesgo En la sección 2-4, introdujimos el concepto general de sesgo. El sesgo puede medirse por medio del **índice de sesgo de Pearson**:

$$I = \frac{3(\bar{x} - \text{mediana})}{s}$$

Si $I \geq 1.00$ o $I \leq -1.00$, los datos se consideran *significativamente sesgados*. Calcule el índice de sesgo de Pearson de los niveles de cotinina de fumadores, listados en la tabla 2-1, y determine si hay un sesgo significativo.

39. Comprensión de la desviación estándar Una muestra consiste en 10 calificaciones de pruebas, que caen entre 70 y 100, inclusive. ¿Cuál es la desviación estándar más grande posible?

40. ¿Datos falsos? Para cualquier conjunto de datos de n valores, con desviación estándar $s\sqrt{n-1}$, cada valor debe estar dentro de la media. Una profesora de estadística reporta que las calificaciones de una prueba que se aplicó a 17 estudiantes de su clase tuvo una media de 75.0 y una desviación estándar de 5.0. Kelly, a quien se considera la mejor estudiante de la clase, afirma haber recibido una calificación de 97. ¿Podría Kelly estar diciendo la verdad?

41. ¿Por qué dividir entre $n - 1$? Sea que una población consista en los valores 3, 6, 9. Suponga que muestras de los valores se seleccionan aleatoriamente *con reemplazo*.

a. Calcule la varianza σ^2 de la población $\{3, 6, 9\}$.

b. Liste las nueve muestras diferentes posibles de los valores seleccionados con reemplazo; luego, calcule la varianza muestral s^2 (que incluye la división entre $n - 1$) de cada una de ellas. Si se seleccionan de forma repetida dos valores muestrales, ¿cuál es el valor medio de la varianza muestral s^2?

c. Para cada una de las nueve muestras, calcule la varianza tratando cada muestra como si fuese una población. (Asegúrese de utilizar la fórmula de la varianza poblacional, que incluye la división entre n). Si selecciona de forma repetida dos valores muestrales, ¿cuál es el valor medio de las varianzas poblacionales?

d. ¿Con qué método se obtienen mejores estimados de σ^2: el inciso *b*) o el inciso *c*)? ¿Por qué? Al calcular las varianzas muestrales, ¿debe utilizarse la división entre n o $n - 1$?

e. Los incisos anteriores muestran que s^2 es un estimador sin sesgo de σ^2. ¿Será s un estimador sin sesgo de σ?

42. ¿Por qué no utilizar la DMA? El ejercicio 41 demuestra que la varianza muestral s^2 es un estimador sin sesgo de σ^2. Haga lo que se le pide con la misma población de $\{3, 6, 9\}$, para demostrar que la desviación media absoluta de una muestra es un estimador sesgado de la desviación media absoluta de una población.

a. Calcule la desviación media absoluta de la población $\{3, 6, 9\}$.

b. Liste las nueve muestras diferentes posibles de dos valores seleccionadas con reemplazo, después calcule la desviación media absoluta de cada una de ellas. Si se seleccionan de forma repetida dos valores muestrales, ¿cuál es el valor medio de las desviaciones medias absolutas?

c. Con base en los resultados de los incisos *a*) y *b*), ¿la desviación media absoluta de una muestra tiende a igualar la desviación media absoluta de una población? ¿La división entre $n - 1$, en lugar de la división entre n, convierte a la desviación media absoluta en un estimado sin sesgo de la desviación media absoluta de la población?

2-6 Medidas de posición relativa

Esta sección incluye medidas que pueden utilizarse para comparar valores de diferentes conjuntos de datos o para comparar valores dentro del mismo conjunto de datos. Aquí introducimos las puntuaciones *z* (para comparar valores de distintos conjuntos de datos), así como los cuartiles y percentiles (para comparar valores dentro del mismo conjunto de datos).

Puntuaciones *z*

Una puntuación *z* (o puntuación estándar) se calcula convirtiendo un valor a una escala estandarizada, como se establece en la siguiente definición. Utilizaremos ampliamente las puntuaciones *z* en el capítulo 5 y en capítulos posteriores, ya que son muy importantes.

Definición

Puntuación estándar o **puntuación z:** número de desviaciones estándar que un valor x se encuentra por arriba o por debajo de la media. Se calcula utilizando las siguientes expresiones:

Muestra $$z = \frac{x - \bar{x}}{s}$$ o *Población* $$z = \frac{x - \mu}{\sigma}$$

(Redondear z a dos espacios decimales).

El siguiente ejemplo ilustra la forma en que se utilizan las puntuaciones z para comparar valores, aun cuando provengan de distintas poblaciones.

EJEMPLO Comparación de estaturas La superestrella de la NBA Michael Jordan mide 78 pulgadas, en tanto que la jugadora de basquetbol de la WNBA Rebecca Lobo mide 76 pulgadas. En efecto, Jordan es más alto por dos pulgadas, pero ¿cuál de los jugadores es *relativamente* más alto? ¿La estatura de Jordan, entre los hombres, excede la estatura de Lobo entre las mujeres? Los hombres tienen estaturas con una media de 69.0 pulgadas, con una desviación estándar de 2.8 pulgadas; las mujeres tienen estaturas con una media de 63.6 pulgadas, con una desviación estándar de 2.5 pulgadas (datos basados en el National Health Survey).

SOLUCIÓN Para comparar las estaturas de Michael Jordan y Rebecca Lobo, en relación con las poblaciones de hombres y mujeres, necesitamos estandarizar dichas estaturas convirtiéndolas en puntuaciones z.

$$\text{Jordan:} \quad z = \frac{x - \mu}{\sigma} = \frac{78 - 69.0}{2.8} = 3.21$$

$$\text{Lobo:} \quad z = \frac{x - \mu}{\sigma} = \frac{76 - 63.6}{2.5} = 4.96$$

INTERPRETACIÓN La estatura de Michael Jordan está a 3.21 desviaciones estándar por arriba de la media, pero la estatura de Rebecca Lobo está a 4.96 desviaciones estándar por arriba de la media. La estatura de Rebecca Lobo entre las mujeres es relativamente mayor que la estatura de Michael Jordan entre los hombres.

Puntuaciones *z* y valores infrecuentes

En la sección 2-5 utilizamos la regla práctica del intervalo para concluir que un valor es "infrecuente" o poco común si está a más de 2 desviaciones estándar de la media. Por lo tanto, los valores infrecuentes tienen puntuaciones z menores que -2 y mayores que 2. (Véase la figura 2-14 en la página 94). Si aplicamos este criterio, tanto Michael Jordan como Rebecca Lobo tienen estaturas infrecuentes, ya que ambos cuentan con estaturas con puntuaciones z mayores que 2.

Si consideramos a jugadores profesionales de basquetbol con estaturas excepcionales, tomemos en cuenta a otro jugador, Mugsy Bogues, que alcanzó el éxito aun cuando sólo mide 5 pies y 3 pulgadas. (Nuevamente usamos el hecho de que

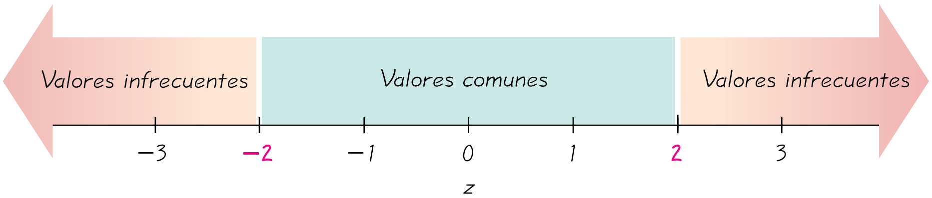


FIGURA 2-14
Interpretación de las puntuaciones *z*
Los valores infrecuentes son aquellos con puntuaciones z menores que -2.00 o mayores que 2.00.

los hombres tienen estaturas con una media de 69.0 pulgadas, con una desviación estándar de 2.8 pulgadas). Después de convertir 5 pies y 3 pulgadas a 63 pulgadas, convertimos su estatura en una puntuación z de la siguiente manera:

$$\text{Bogues: } z = \frac{x - \mu}{\sigma} = \frac{63 - 69.0}{2.8} = -2.14$$

Demos gracias a Mugsy Bogues por sus muchos años de juego inspirado y por ilustrar este principio:

Siempre que un valor sea menor que la media, su puntuación *z* correspondiente será negativa

Valores comunes: $\mathbf{-2 \leq z}$ **puntuación** $\mathbf{\leq 2}$

Valores infrecuentes: $\mathbf{z}$ **puntuación** $\mathbf{< -2}$ *o* $\mathbf{z}$ **puntuación** $\mathbf{> 2}$

Las puntuaciones z son medidas de posición, en el sentido de que describen la localización de un valor (en términos de desviaciones estándar), en relación con la media. Una puntuación z de 2 indica que un valor está a dos desviaciones estándar *por encima* de la media, en tanto que una puntuación z de -3 indica que un valor está a tres desviaciones estándar *por debajo* de la media. Los cuartiles y los percentiles también son medidas de posición, pero se definen de forma distinta que las puntuaciones z; son útiles para comparar valores dentro del mismo conjunto de datos o entre distintos conjuntos de datos.

Cuartiles y percentiles

De la sección 2-4, recuerde que la mediana de un conjunto de datos es el valor que está a la mitad, de modo que 50% de los valores son iguales o menores a la mediana y el 50% de los valores son mayores o iguales a la mediana. Tal como la mediana divide los datos en dos partes iguales, los tres **cuartiles**, denotados por Q_1, Q_2 y Q_3, dividen los valores ordenados en cuatro partes iguales. (Los valores están *ordenados* cuando se acomodan en orden).

He aquí descripciones de los tres cuartiles:

Q_1 (Primer cuartil): Separa el 25% inferior de los valores ordenados, del 75% superior. (Para ser más precisos, al menos el 25% de los valores ordenados son menores o iguales que Q_1, y al menos el 75% de los valores son mayores o iguales que Q_1).

Q_2 (Segundo cuartil): Igual a la mediana; separa el 50% inferior de los valores ordenados, del 50% superior.

Q_3 (Tercer cuartil): Separa el 75% inferior de los valores ordenados, del 25% superior. (Para ser más precisos, al menos el 75% de los valores ordenados son menores o iguales que Q_3, y al menos el 25% de los valores son mayores o iguales que Q_3).

Describiremos un procedimiento para el cálculo de cuartiles después de analizar los percentiles. No existe un acuerdo universal respecto de un procedimiento único para el cálculo de cuartiles, y con frecuencia los distintos programas de cómputo producen resultados diferentes. Por ejemplo, si usted utiliza un conjunto de datos 1, 3, 6, 10, 15, 21, 28 y 36, obtendrá los siguientes resultados:

	Q_1	Q_2	Q_3
STATDISK	4.5	12.5	24.5
Minitab	3.75	12.5	26.25
Excel	5.25	12.5	22.75
TI-83 Plus	4.5	12.5	24.5

Para este conjunto de datos, STATDISK y la calculadora TI-83 Plus coinciden, aunque esto no siempre sucede. Si utiliza una calculadora o un programa de cómputo para resolver ejercicios que comprenden cuartiles es posible que obtenga resultados que difieran ligeramente de las respuestas que vienen al final del libro.

Así como hay tres cuartiles que separan un conjunto de datos en cuatro partes, también se tienen 99 **percentiles**, que se denotan $P_1, P_2, \ldots, P_{99}$, los cuales separan los datos en 100 grupos, con aproximadamente el 1% de los valores en cada grupo. (Los cuartiles y percentiles son ejemplos de *cuantiles* o *fractiles*, que separan los datos en grupos con casi el mismo número de valores).

El proceso para calcular percentiles, que corresponde a un valor particular x, es bastante sencillo, tal como se indica en la siguiente expresión:

$$\text{percentil del valor } x = \frac{\text{número de valores menores que } x}{\text{número total de valores}} \cdot 100$$

EJEMPLO Niveles de cotinina de fumadores La tabla 2-13 lista los 40 niveles ordenados de cotinina de fumadores que se incluyen en la tabla 2-1. Calcule el percentil correspondiente al nivel de cotinina de 112.

SOLUCIÓN A partir de la tabla 2-13 se ve que hay dos valores menores que 112; por lo tanto,

$$\text{percentil de } 112 = \frac{12}{40} \cdot 100 = 30$$

INTERPRETACIÓN El nivel de cotinina de 112 es el percentil 30o.

El ejemplo anterior muestra cómo convertir un valor muestral dado a su percentil correspondiente. Existen diversos métodos para el procedimiento inverso de convertir un percentil en el valor correspondiente del conjunto de datos. El procedimiento que usaremos se resume en la figura 2-15, que emplea la notación que viene inmediatamente después.

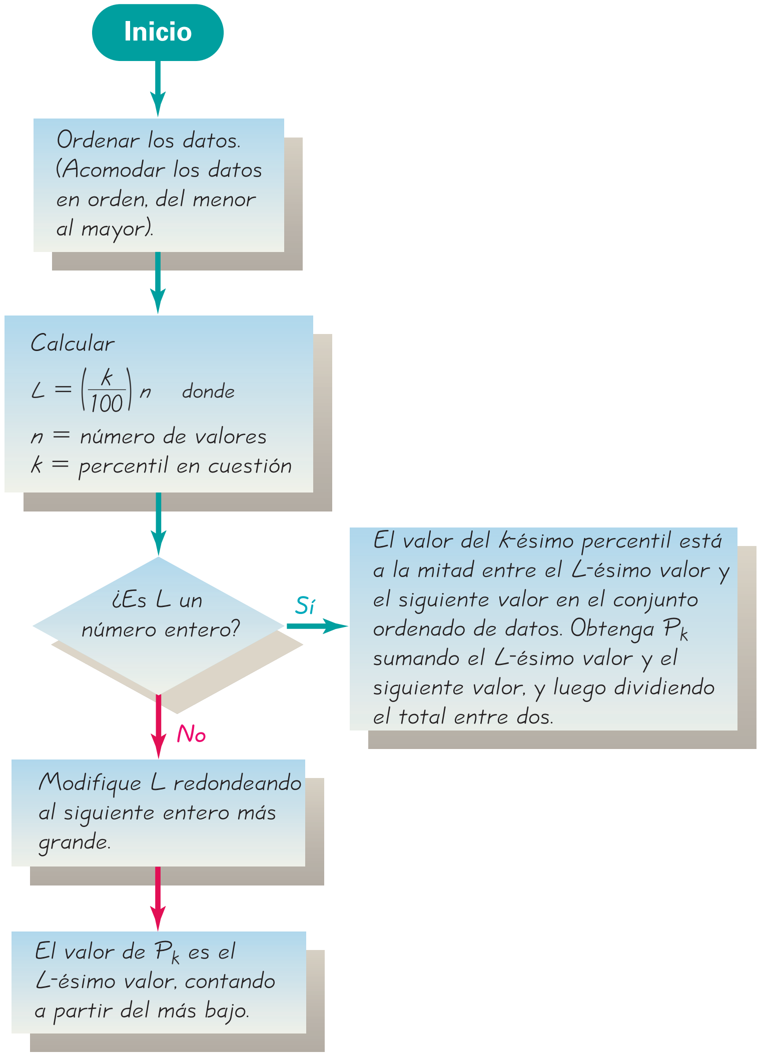
Tabla 2-13 *Niveles ordenados de cotinina de 40 fumadores*

0	1	1	3	17	32	35	44	48	86
87	103	112	121	123	130	131	149	164	167
173	173	198	208	210	222	227	234	245	250
253	265	266	277	284	289	290	313	477	491

Índice del costo de la risa

En realidad hay un Índice del Costo de la Risa (ICR), que busca los costos de artículos como pollos de plástico, anteojos de Groucho Marx, entradas a clubes de comediantes y otros 13 indicadores principales del humor. Éste es el mismo método básico que se utiliza en la creación del Índice de Precios al Consumidor (IPC), que se basa en un promedio ponderado de bienes y servicios adquiridos por consumidores comunes. Mientras que las puntuaciones estándar y los percentiles nos permiten comparar valores diferentes, ignorando cualquier elemento del tiempo, los números índice, tales como el ICR y el IPC, nos permiten comparar el valor de alguna variable con su valor en un periodo base. El valor de un número índice es el valor real dividido entre el valor base, multiplicado por 100.

FIGURA 2-15 **Conversión del *k*-ésimo percentil al valor del dato correspondiente**



Notación

n = número total de valores en el conjunto de datos
k = percentil utilizado (ejemplo: para el percentil 25o, $k = 25$).
L = localizador que da la *posición* de un valor (ejemplo: para el valor 12o en la lista ordenada, $L = 12$).
P_k = percentil k-ésimo (ejemplo: P_{25} es el percentil 25o).

EJEMPLO Niveles de cotinina de fumadores Remítase a los niveles ordenados de cotinina de fumadores de la tabla 2-13, y utilice la figura 2-15 para calcular el valor del percentil 68o, P_{68}.

SOLUCIÓN Si nos referimos a la figura 2-15, veremos que los datos muestrales ya están ordenados, de modo que es posible proceder al cálculo del valor del localizador L. En el cálculo utilizamos $k = 68$, ya que estamos tratando de obtener el valor del percentil 68o. Usamos $n = 40$, porque tenemos 40 valores de datos.

$$L = \frac{k}{100} \cdot n = \frac{68}{100} \cdot 40 = 27.2$$

Después, nos preguntamos si L es un número entero; respondemos que no. Por lo tanto, procederemos al siguiente recuadro hacia abajo y modificamos L, redondeándola de 27.2 a 28. (En este libro solemos redondear de la forma común, pero es uno de los casos donde redondeamos hacia *arriba* y no hacia el entero *más cercano*). Por último, el recuadro final muestra que el valor de P_{68} es el valor 28o, contando hacia arriba, desde el mínimo. En la tabla 2-13, el valor 28o es 234. Es decir, $P_{68} = 234$.

PROBLEMA DEL CAPÍTULO 2

EJEMPLO Niveles de cotinina de fumadores Remítase a la muestra de niveles de cotinina de fumadores que aparece en la tabla 2-13. Utilice la figura 2-15 para calcular el valor de Q_1, que es el primer cuartil.

SOLUCIÓN Primero observamos que Q_1 es igual que P_{25}, por lo que procedemos a calcular el valor del percentil 25o. Si nos referimos a la figura 2-15, veremos que los datos muestrales ya se ordenaron, de manera que procedemos a calcular el valor del localizador L. En este cálculo utilizamos $k = 25$, ya que tratamos de obtener el valor del percentil 25o y usamos $n = 40$, porque tenemos 40 valores de datos.

$$L = \frac{k}{100} \cdot n = \frac{25}{100} \cdot 40 = 10$$

Después, nos preguntamos si L es un número entero; respondemos que sí. Por lo tanto, vamos al recuadro que se ubica a la derecha. Vemos que el valor del percentil k-ésimo (25o) está a la mitad entre el valor L-ésimo (10o) y el siguiente valor en el conjunto original de datos. Es decir, el valor del percentil 25o se ubica a la mitad, entre el 10o valor y el 11o valor. El 10o valor es 86 y el 11o valor es 87; por lo tanto, el valor a la mitad de ellos es 86.5. Concluimos que el percentil 25o es $P_{25} = 86.5$. El valor del primer cuartil Q_1 es también 86.5.

El ejemplo anterior demuestra que al calcular un valor cuartilar (como Q_1), es posible utilizar el valor del percentil equivalente (como P_{25}) en su lugar. Al margen, se indican las relaciones equivalentes entre cuartiles y percentiles.

$Q_1 = P_{25}$
$Q_2 = P_{50}$
$Q_3 = P_{75}$

En secciones anteriores de este capítulo describimos diversos estadísticos, incluyendo media, mediana, moda, rango y desviación estándar. Algunos otros estadísticos se definen con el uso de cuartiles y percentiles, como los siguientes:

$$\text{rango intercuartilar (o RIC)} = Q_3 - Q_1$$

$$\text{rango semiintercuartilar} = \frac{Q_3 - Q_1}{2}$$

$$\text{cuartil medio} = \frac{Q_3 + Q_1}{2}$$

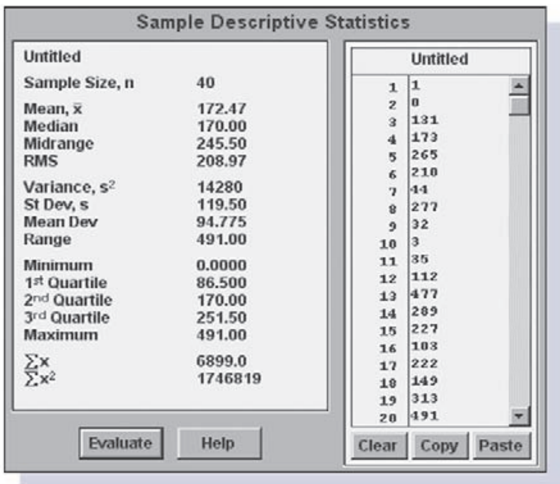
$$\text{rango de percentiles 10—90} = P_{90} - P_{10}$$

Después de completar esta sección, usted debe ser capaz de convertir un valor en su puntuación z (o puntuación estándar) correspondiente, de manera que sea posible compararlo con otros valores que provienen de diferentes conjuntos de datos. También tendrá que ser capaz de convertir un valor en su valor percentil correspondiente, de manera que pueda compararlo con otros valores en algún conjunto de datos. También sabrá convertir un percentil en su valor de dato correspondiente. Finalmente, comprenderá el significado de los cuartiles y podrá relacionarlos con sus valores percentiles correspondientes (como en $Q_3 = P_{75}$).

Utilizando la tecnología

Se puede utilizar una variedad de programas de cómputo y calculadoras diferentes para calcular muchos de los estadísticos estudiados hasta ahora en este capítulo. En la sección 2-4 dimos instrucciones específicas para el uso de STATDISK, Minitab, Excel y la calculadora TI-83 Plus. Señalamos que en ocasiones es posible introducir un conjunto de datos y utilizar una operación para obtener diversos estadísticos muestrales, frecuentemente llamados *estadísticos descriptivos*. En las siguientes representaciones visuales se mencionan ejemplos de tales resultados, los cuales provienen de los niveles de cotinina de *fumadores* que se presentaron en la tabla 2-1, en el problema del capítulo. Los resultados de la calculadora TI-83 Plus se muestran en dos pantallas, ya que no caben en una sola.

STATDISK



Minitab

Variable	N	Mean	Median	TrMean	StDev	SE Mean
SMOKER	40	172.5	170.0	164.7	119.5	18.9

Variable	Minimum	Maximum	Q1	Q3
SMOKER	0.0	491.0	86.3	252.3

Excel

Column1	
Mean	172.475
Standard Error	18.89434
Median	170
Mode	1
Standard Deviation	119.4983
Sample Variance	14279.85
Kurtosis	0.519621
Skewness	0.587929
Range	491
Minimum	0
Maximum	491
Sum	6899
Count	40

TI-83 Plus

```
1-Var Stats
 x̄=172.475
 Σx=6899
 Σx²=1746819
 Sx=119.4983076
 σx=117.9951244
↓n=40
```

```
1-Var Stats
↑n=40
 minX=0
 Q₁=86.5
 Med=170
 Q₃=251.5
 maxX=491
```

2-6 Destrezas y conceptos básicos

En los ejercicios 1 a 4 exprese todas las puntuaciones z con dos decimales.

1. Puntuaciones de CI Las puntuaciones de CI de la prueba Stanford Binet tienen una media de 100 y una desviación estándar de 16. Albert Einstein obtuvo un CI de 160.
a. ¿Cuál es la diferencia entre el CI de Einstein y la media?
b. ¿Cuántas desviaciones estándar implica esto [la diferencia obtenida en el inciso *a*)]?
c. Convierta la puntuación de CI de Einstein a puntuación *z*.
d. Si consideramos que las puntuaciones de CI "comunes" son aquellas que, convertidas en puntuaciones *z*, caen entre -2 y 2, ¿es el CI de Einstein común o infrecuente?

2. Pulso de adultos Suponga que los adultos tienen pulsos (latidos por minuto) con una media de 72.9 y una desviación estándar de 12.3 (con base en datos del National Health Examination). Cuando escribió este ejercicio, el autor tenía un pulso de 48.
a. ¿Cuál es la diferencia entre el pulso del autor y la media?
b. ¿Cuántas desviaciones estándar implica esto [la diferencia obtenida en el inciso *a*)]?
c. Convierta el pulso de 48 a puntuación *z*.
d. Si consideramos que los pulsos "comunes" son aquellos que, convertidos en puntuaciones *z*, caen entre -2 y 2, ¿es el pulso de 48 común o infrecuente? ¿Podría explicar por qué un pulso sería inusualmente bajo? (La razón de este pulso tan bajo *no* es que los autores de libros de estadística estén comúnmente en un estado que se describiría como comatoso).

3. Estaturas de hombres Las estaturas de hombres adultos tienen una media de 69.0 pulgadas y una desviación estándar de 2.8 pulgadas. Calcule las puntuaciones *z* que corresponden a los siguientes individuos:
a. El actor Danny DeVito, que mide 5 pies.
b. El jugador de basquetbol de la NBA, Shaquille O'Neal, que mide 7 pies 1 pulgada.
c. El autor, quien es un "jugador" de golf y tenis, que mide 69.72 pulgadas.

4. Temperaturas corporales La temperatura corporal humana tiene una media de 98.20°F y una desviación estándar de 0.62°F. Convierta las temperaturas dadas a puntuaciones z.
a. 100° **b.** 96.96° **c.** 98.20°

En los ejercicios 5 a 8 exprese todas las puntuaciones z con dos decimales. Considere una puntuación como infrecuente si es menor que −2.00 o mayor que 2.00.

5. Estaturas de mujeres El Club Beanstalk es sólo para mujeres y hombres muy altos. La estatura mínima que se requiere en las mujeres es de 70 pulgadas. Las estaturas de las mujeres tienen una media de 63.6 pulgadas y una desviación estándar de 2.5 pulgadas. Calcule la puntuación z correspondiente a una mujer con una estatura de 70 pulgadas; después, determine si dicha estatura es infrecuente.

6. Duración del embarazo Una mujer, que escribió a *Dear Abby*, afirmó que dio a luz 308 días después de una visita de su esposo, que estaba en la marina. La duración del embarazo tiene una media de 268 días y una desviación estándar de 15 días. Calcule la puntuación z de 308 días. ¿Es infrecuente la duración? ¿Qué concluye?

7. Temperatura corporal La temperatura corporal humana tiene una media de 98.20°F y una desviación estándar de 0.62°F. Se descubre que un paciente de urgencias tiene una temperatura de 101°F. Convierta 101° en puntuación z. ¿Es la temperatura inusualmente alta? ¿Qué sugiere esto?

8. Niveles de colesterol Para hombres de entre 18 y 24 años de edad, los niveles séricos de colesterol (en mg/100ml) tienen una media de 178.1 y una desviación estándar de 40.7 (con base en datos del National Health Survey). Calcule la puntuación z correspondiente de un hombre de entre 18 y 24 años, quien presenta un nivel sérico de colesterol de 259.0 mg/100ml. ¿Es este nivel inusualmente alto?

9. Comparación de calificaciones de una prueba ¿Cuál es relativamente mejor: una calificación de 85 en una prueba de psicología o una calificación de 45 en una prueba de economía? Las calificaciones en la prueba de psicología tienen una media de 90 y una desviación estándar de 10. Las calificaciones en la prueba de economía tienen una media de 55 y una desviación estándar de 5.

10. Comparación de calificaciones Tres estudiantes resuelven pruebas equivalentes del sentido del humor; una vez que la risa disminuye, se calculan sus calificaciones. ¿Cuál es la calificación relativa más alta?
a. Una calificación de 144 en una prueba que tiene una media de 128 y una desviación estándar de 34.
b. Una calificación de 90 en una prueba que tiene una media de 86 y una desviación estándar de 18.
c. Una calificación de 18 en una prueba que tiene una media de 15 y una desviación estándar de 5.

T 11. Peso de Coca Cola Remítase en el conjunto de datos 17 del Apéndice B, a la muestra de 36 pesos de Coca Cola regular. Convierta en peso de 0.7901 en puntuación z. ¿Es 0.7901 un peso inusual de la Coca Cola regular?

T 12. M&M verdes Remítase en el conjunto de datos 19 del apéndice B, a la muestra de pesos de dulces M&M verdes. Convierta el peso del M&M verde más pesado en puntuación z. ¿Es infrecuente el peso del M&M verde más pesado en estos dulces?

En los ejercicios 13 a 16 utilice los 40 niveles ordenados de cotinina de fumadores, listados en la tabla 2.13. Calcule el percentil correspondiente a los niveles de cotinina dados.

13. 149 14. 210 15. 35 16. 250

En los ejercicios 17 a 24 utilice los 40 niveles ordenados de cotinina de fumadores, listados en la tabla 2.13. Calcule el percentil o cuartil indicado.

17. P_{20} 18. Q_3 19. P_{75} 20. Q_2
21. P_{33} 22. P_{21} 23. P_1 24. P_{85}

En los ejercicios 25 a 28 utilice los niveles de colesterol de mujeres listados en el conjunto de datos 1 del Apéndice B. Calcule el percentil correspondiente al nivel de colesterol dado.

25. 123 **26.** 309 **27.** 271 **28.** 126

En los ejercicios 29 a 36 utilice los niveles de colesterol de mujeres listados en el conjunto de datos 1 del Apéndice B. Calcule el percentil o cuartil indicado.

29. P_{85} **30.** P_{35} **31.** Q_1 **32.** Q_3

33. P_{18} **34.** P_{36} **35.** P_{58} **36.** P_{96}

2-6 Más allá de lo básico

37. Unidades de medición Cuando se calcula una puntuación z para la estatura de un jugador de basquetbol de la NBA, ¿de qué manera se afecta el resultado si, en lugar de utilizar pulgadas, todas las estaturas se expresan en centímetros? En general, ¿de qué manera se afectan las puntuaciones z por la unidad particular de medición que se utiliza?

38. Conversión de una puntuación z La estatura de las mujeres tiene una media de 63.6 pulgadas y una desviación estándar de 2.5 pulgadas.

a. Julia Roberts, que es una de las actrices más exitosas de los últimos años, tiene una estatura que, convertida a puntuación z, *es* de 2.16. ¿Qué tan alta es (en pulgadas)?

b. La cantante de rap Lil'Kim tiene una estatura que, convertida a puntuación z, *es* de -1.84. ¿Qué tan alta es (en pulgadas)?

39. Distribución de puntuaciones z

a. Un conjunto de datos tiene una distribución uniforme. Si todos los valores se convierten a puntuaciones z, ¿cuál es la forma de la distribución de las puntuaciones z?

b. Un conjunto de datos tiene una distribución normal. Si todos los valores se convierten a puntuaciones z, ¿cuál es la forma de la distribución de las puntuaciones z?

c. En general, ¿cómo se ve afectada la forma de una distribución si todos los valores se convierten en puntuaciones z?

40. Secuencia de Fibonacci Éstos son los primeros de muchos términos de la famosa secuencia de Fibonacci: 1, 1, 2, 3, 5, 8, 13.

a. Calcule la media $\bar{x}$ y la desviación estándar s; después, convierta cada valor a puntuación z. No redondee las puntuaciones z. Ocupe tantos datos como su calculadora pueda manejar.

b. Calcule la media y la desviación estándar de las puntuaciones z que se obtuvieron en el inciso *a*).

c. Si utilizara cualquier otro conjunto de datos, ¿obtendría los mismos resultados que en el inciso *b*)?

41. Niveles de cotinina de fumadores Utilice los niveles ordenados de cotinina de fumadores que se listan en la tabla 2-3.

a. Calcule el rango intercuartilar.

b. Calcule el cuartil medio.

c. Calcule el rango de percentiles 10-90.

d. ¿Es $P_{50} = Q_2$? Si es así, ¿es P_{50} *siempre* igual a Q_2?

e. ¿Es $Q_2 = (Q_1 + Q_3)/2$? Si es así, ¿es Q_2 *siempre* igual a $(Q_1 + Q_3)/2$?

42. Interpolación Cuando se calculan percentiles con el uso de la figura 2-15, si el localizador L no es un número entero, lo redondeamos hacia el siguiente número entero mayor. Una alternativa para este procedimiento es interpolar, de modo que un localizador de 23.75 conduce a un valor que está a 0.75 (o 3/4) del camino entre los valores 23o y 24o. Utilice este método de interpolación para calcular P_{35} y Q_1 para los pesos de los osos que se listan en el conjunto de datos 9 del Apéndice B.

43. Deciles y cuartiles En un conjunto de datos hay nueve **deciles**, que se denotan con D_1, $D_2, \ldots, D_9$, que dividen los datos ordenados en 10 grupos, con aproximadamente 10% de los valores en cada grupo. También existen cuatro **quintiles**, que dividen los datos ordenados en cinco grupos, con aproximadamente 20% de los valores en cada grupo. (Note la diferencia entre los quintiles y los cuartiles, que ya describimos en esta sección).

a. ¿Qué percentil es equivalente a D_1? ¿A D_5? ¿A D_8?

b. Utilice los niveles ordenados de cotinina de fumadores de la tabla 2-13 y calcule los nueve deciles.

c. Utilice los niveles ordenados de cotinina de fumadores de la tabla 2-13 y calcule los cuatro quintiles.

2-7 Análisis exploratorio de datos (AED)

El presente capítulo presenta las herramientas básicas para describir, explorar y comparar datos; esta sección se enfoca en la exploración de datos. Iniciamos definiendo el análisis exploratorio de datos; después, introduciremos los datos distantes, el resumen de 5 números y las gráficas de cuadro.

Definición

Análisis exploratorio de datos: proceso para utilizar herramientas estadísticas (como gráficas, medidas de tendencia central y medidas de variación), con la finalidad de investigar conjuntos de datos para comprender sus características importantes.

Recuerde que en la sección 2-1 mencionamos cinco características importantes de los datos, y que iniciamos con **1.** *el centro*, **2.** *la variación*, y **3.** la naturaleza de la *distribución*. Tales características pueden investigarse calculando los valores de la media y la desviación estándar, así como por medio de la construcción de un histograma. Por lo general, es importante investigar más el conjunto de datos para identificar cualquier particularidad notable, en especial aquélla que llegue a afectar de forma importante los resultados y las conclusiones. Una de estas características es la presencia de datos distantes.

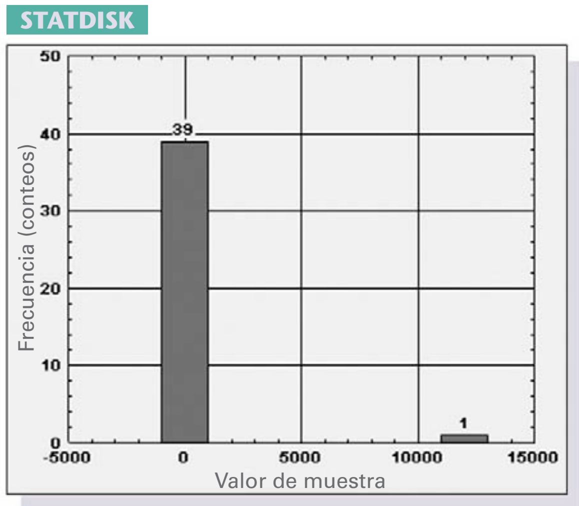
Datos distantes

Dato distante: valor que está muy alejado de la mayoría de los demás valores. Un dato distante es un valor *extremo* en relación con los otros datos. Cuando se explora un conjunto de datos, se deben considerar los datos distantes, ya que pueden revelar información importante y afectar, en gran medida, el valor de la media y de la desviación estándar, así como distorsionar gravemente un histograma. El siguiente ejemplo utiliza un valor incorrecto para un dato distante; aunque no todos los datos distantes son errores, algunos de ellos son valores correctos.

PROBLEMA 2 DEL CAPÍTULO

EJEMPLO Niveles de cotinina de fumadores Cuando se utiliza un programa de cómputo o una calculadora, es muy fácil cometer errores con los dedos. Remítase a los niveles de cotinina de fumadores que se listan en la tabla 2-1, en el problema del capítulo; suponga que el primer dato de 1 se introduce de manera incorrecta como 11111, porque usted estaba distraído viendo un meteorito que aterrizaba en su jardín. El dato incorrecto de 11111 es un dato distante, ya que se localiza muy lejos de los demás valores. ¿De qué manera afecta ese dato distante a la media, a la desviación estándar y al histograma?

SOLUCIÓN Cuando el dato 1 se reemplaza con el valor distante de 11111, la media cambia de 172.5 a 450.2, de modo que el efecto del dato distante es muy grande. El dato incorrecto de 11111 causa que la desviación estándar cambie de 119.5 a 1732.7, por lo que el efecto del dato distante también es muy grande. La figura 2-1, en la sección 2-3, muestra el histograma con los valores correctos de los niveles de cotinina de fumadores de la tabla 2-1, pero la representación visual del STATDISK que se muestra aquí, contiene el histograma que resulta del uso de los mismos datos con el valor de 1, reemplazado por el valor incorrecto de 11111. Compare el histograma del STATDISK con la figura 2-1 y verá fácilmente que la presencia del dato distante afecta de manera drástica la forma de la distribución.



El ejemplo anterior ilustra estos principios importantes:

1. **Un dato distante puede tener un efecto importante sobre la media.**
2. **Un dato distante puede tener un efecto importante sobre la desviación estándar.**
3. **Un dato distante puede tener un efecto importante sobre la escala del histograma, de modo que la verdadera naturaleza de la distribución se oculta totalmente.**

Un procedimiento sencillo para encontrar datos distantes es el examen de una lista *ordenada* de los datos. En particular, observe los valores mínimo y máximo muestrales; luego, determine si se alejan mucho de los demás valores. Algunos datos distantes son valores correctos y algunos son errores, como en el ejemplo anterior. Si estamos seguros de que un dato distante es un error, debemos corregirlo o eliminarlo. Si incluimos un dato distante, porque sabemos que es correcto, podríamos estudiar sus efectos por medio de la construcción de gráficas y el cálculo de estadísticos que incluyan y que no incluyan los datos distantes.

Una propina extrema

Es importante tomar en cuenta los datos distantes ya que, en muchos casos, un valor extremo puede tener un efecto muy importante en los estadísticos y en las conclusiones que se derivan de ellos. A veces un dato distante es un error que debe ser corregido o eliminado. En otros, un dato distante es un valor válido que debe investigarse para obtener información importante. Algunos alumnos del autor, al reunir datos consistentes de facturas y propinas de restaurantes, no encontraron datos distantes sobresalientes en esos valores muestrales. Sin embargo, un dato distante es la propina de 16,000 dólares que se dio por una cuenta de 8,899.78 dólares en un restaurante. Esta propina la dio un ejecutivo de Londres no identificado al mesero Lenny Lorando, en el restaurante Nello's, ubicado en la ciudad de Nueva York. Lorando dijo que ya antes había atendido al cliente y que "él siempre es generoso, pero nunca antes de esta forma. Tengo que hablarle a mi hermana acerca de él".

LA ESTADÍSTICA EN LAS NOTICIAS

Buen consejo para los periodistas

El columnista Max Frankel escribió en el *New York Times* que "la mayoría de las escuelas de periodismo dan poca importancia a la estadística y algunas permiten que los estudiantes se gradúen sin entrenamiento alguno en números. ¿Cómo pueden estos reporteros escribir con sensibilidad sobre el comercio, la asistencia social y el crimen, o sobre tarifas aéreas, la atención a la salud y la nutrición? El uso sentimental que hacen los medios de comunicación de los números acerca de la incidencia de accidentes o muertes atemoriza a las personas y las deja vulnerables a las exageraciones periodísticas, la demagogia política y el fraude comercial". Este escritor cita varios casos, incluyendo el ejemplo de un artículo de página completa acerca del déficit de la ciudad de Nueva York con la promesa del alcalde de cubrir el déficit presupuestal de 2.7 mil millones de dólares; en el artículo nunca se menciona el tamaño *total* del presupuesto, de modo que la cifra de 2.7 mil millones de dólares está fuera de contexto.

Gráficas de cuadro

Además de las gráficas presentadas en la sección 2.3, una gráfica de cuadro es otro tipo de gráfica que se utiliza a menudo. Las gráficas de cuadro son útiles para revelar la tendencia central de los datos, su dispersión, su distribución y la presencia de datos distantes. La construcción de una gráfica de cuadro requiere que primero se obtenga el valor mínimo, el valor máximo y los cuartiles, tal como se define en el resumen de los cinco números.

Definiciones

Para un conjunto de datos, el **resumen de los cinco números** consiste en el valor mínimo; el primer cuartil, Q_1; la mediana (o segundo cuartil, Q_2) ; el tercer cuartil, Q_3; y el valor máximo.

Gráfica de cuadro (o **diagrama de cuadro y bigotes**): gráfica de un conjunto de datos que consiste en una línea que se extiende desde el valor mínimo hasta el valor máximo, así como una caja con líneas trazadas en el primer cuartil, Q_1; la mediana y el tercer cuartil, Q_3. (Véase la figura 2-16).

Procedimiento para construir una gráfica de cuadro

1. Elabore el resumen de los cinco números, consistente en el valor mínimo, Q_1, la mediana, Q_3, y el valor máximo.
2. Construya una escala con valores que incluyan el valor mínimo y el valor máximo.
3. Construya un cuadro (un rectángulo) que se extienda desde Q_1 hasta Q_3, y dibuje una línea en la caja, en el valor de la mediana.
4. Dibuje líneas que se extiendan hacia afuera del cuadro, hasta los valores mínimo y máximo.

Las gráficas de cuadro no muestran tanta información detallada como los histogramas o las gráficas de tallo y hojas, por lo que podría no ser la mejor elección cuando se maneja un solo conjunto de datos. Suelen ser muy útiles para comparar dos o más conjuntos de datos. Cuando se utilicen dos o más gráficas de cuadro para comparar distintos conjuntos de datos, es importante emplear la misma escala, de manera que sea posible realizar comparaciones correctas.

PROBLEMA DEL CAPÍTULO 2

EJEMPLO Niveles de cotinina de fumadores Remítase a los 40 niveles de cotinina de fumadores en la tabla 2-1 (sin el error de 11111 utilizado en lugar del 1 en el ejemplo anterior).

a. Obtenga los valores que constituyen el resumen de los cinco números.

b. Construya una gráfica de cuadro.

SOLUCIÓN

a. El resumen de los cinco números consta del valor mínimo, Q_1, la mediana, Q_3, y el valor máximo. Para obtener dichos valores, primero ordene los datos (acomódelos en orden del más bajo al más alto). El mínimo de 0 y el máximo de 491 son fáciles de identificar en la lista ordenada. Ahora, proceda a calcular los cuartiles. Si usamos el diagrama de flujo de la figura 2-15, obtendremos $Q_1 = P_{25} = 86.5$, que se sitúa al calcular el localizador $L = (25/100)40 = 10$, y al encontrar el valor que está a la mitad entre el 10o y el 11o valores en la lista ordenada. La mediana es 170, que es el valor que está a la mitad entre los valores 20o y 21o. También encontramos que $Q_3 = 251.5$, al utilizar la figura 2.15 para el percentil 75o. Por lo tanto, el resumen de los cinco números es 0, 86.5, 170, 251.5 y 491.

b. En la figura 2-16 creamos la gráfica de cuadro para los datos. Usamos el valor mínimo (0) y el valor máximo (491) para determinar la escala de valores; después, graficamos los valores del resumen de los cinco números, como se indica a continuación.

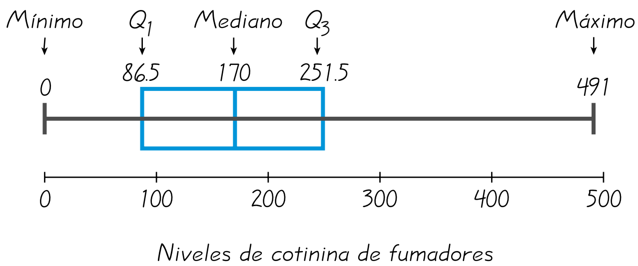


FIGURA 2-16

Gráfica de cuadro

En la figura 2-17, presentamos algunas gráficas de cuadro genéricas, junto con formas comunes de distribución. Parece ser que los niveles de cotinina de fumadores tienen una distribución sesgada.

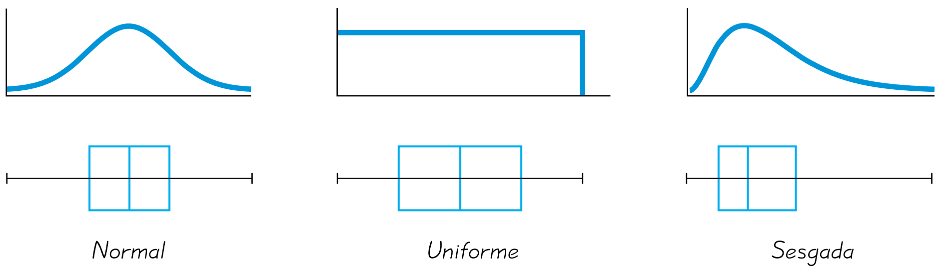
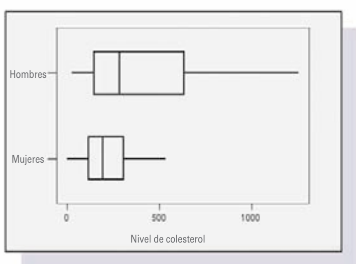


FIGURA 2-17 Gráficas de cuadro que corresponden a distribuciones normales, uniformes y sesgadas

Para ilustrar el uso de gráficas de cuadro que permiten comparar conjuntos de datos, véase la representación visual de Minitab de los niveles de colesterol para una muestra de hombres y una muestra de mujeres, con base en datos del National Health Examination, que se incluyen en el conjunto de datos 1 del Apéndice B. De acuerdo con el conjunto de datos, parece que los hombres tienen niveles de colesterol generalmente más altos que las mujeres, y que los niveles de colesterol de los hombres varían más que los de las mujeres.



LA ESTADÍSTICA EN LAS NOTICIAS

"Mejores" universidades

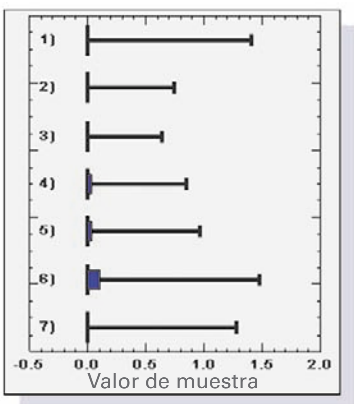
Cada año, el *U.S. News and World Report* publica un número con una lista de "las mejores universidades de Estados Unidos". Generalmente las ventas de ese ejemplar aumentan hasta un 40%. Existen críticos de la lista que argumentan en contra de los criterios y el método de recolección de datos. Las quejas más comunes son: que se da demasiada importancia a los criterios de riqueza, la reputación, las calificaciones del consejo universitario, las donaciones de los alumnos y las opiniones de los presidentes universitarios; que se da poca importancia a la satisfacción de los estudiantes y a las prácticas educativas efectivas. El *New York Times* entrevistó a Kenneth Auchincloss, que es editor de la obra *How to Get into College* (de Kaplan/Newsweek), quien respondió que "nunca nos hemos sentido cómodos tratando de cuantificar en términos numéricos los diversos criterios empleados al calificar a una universidad como buena o menos buena, y no queremos dedicar los recursos a realizar un análisis estadístico elaborado que, con franqueza, no pensamos que sea válido".

EJEMPLO ¿Llueve más durante los fines de semana? Remítase al conjunto de datos 11 del Apéndice B, que incluye una lista de las cantidades de lluvia (en pulgadas) que cayeron en Boston todos los días de un año reciente. La reunión de este conjunto de datos se inspiró con reportes de los medios de comunicación acerca de que llueve más durante los fines de semana (sábado y domingo) que entre semana. Más adelante, en este libro, describiremos métodos estadísticos importantes que permitan probar, de manera formal, dicha aseveración; por ahora, exploremos el conjunto de datos, para ver qué puede aprenderse. (Aun cuando sepamos aplicar estos métodos estadísticos formales, primero habrá que explorar los datos, antes de proceder con el análisis formal.)

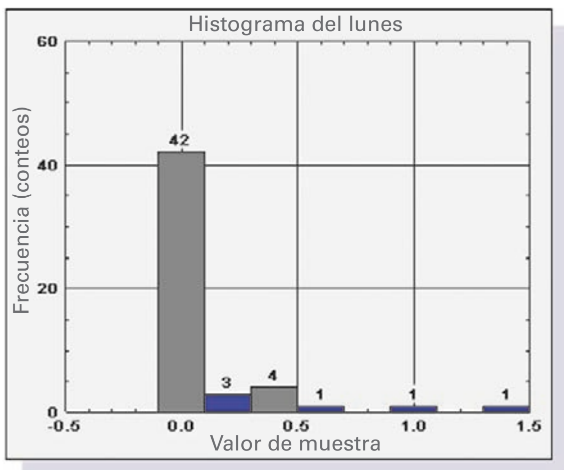
SOLUCIÓN Comencemos investigando los elementos clave del centro, la variación, la distribución, los datos distantes y las características en el tiempo (la misma lista "CVDDT" que se introdujo en la sección 2-1). Abajo se presentan medidas de tendencia central (media), medidas de variación (desviación estándar) y el resumen de los cinco números para las cantidades de lluvia que caen cada día de la semana. La representación visual del STATDISK muestra gráficas de cuadro de cada uno de los siete días de la semana, iniciando con el lunes en la parte superior. Debido a que los histogramas de los siete días son muy similares, únicamente mostramos el histograma de las cantidades de lluvia del lunes.

	Media	Desviación estándar	Mínimo	Q_1	Mediana	Q_3	Máximo
Lunes	0.100	0.263	0.000	0.000	0.000	0.010	1.410
Martes	0.058	0.157	0.000	0.000	0.000	0.015	0.740
Miércoles	0.051	0.135	0.000	0.000	0.000	0.010	0.640
Jueves	0.069	0.167	0.000	0.000	0.000	0.040	0.850
Viernes	0.095	0.228	0.000	0.000	0.000	0.040	0.960
Sábado	0.143	0.290	0.000	0.000	0.000	0.100	1.480
Domingo	0.068	0.200	0.000	0.000	0.000	0.010	1.280

STATDISK



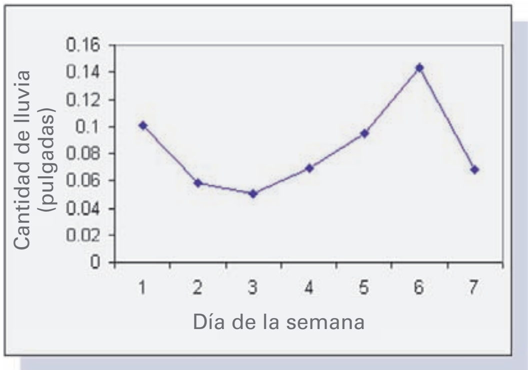
STATDISK



INTERPRETACIÓN Al examinar y comparar los estadísticos y las gráficas, hicimos las siguientes observaciones importantes:

- *Medias*: Las medias varían desde un mínimo de 0.051 pulgadas hasta un máximo de 0.143 pulgadas. Las siete medias varían en cantidades considerables. En capítulos siguientes presentaremos métodos para determinar si tales diferencias son *significativas*. (Métodos posteriores mostrarán que las medias no difieren en cantidades significativas). Si colocamos las medias en orden de menor a mayor, obtendremos la siguiente secuencia de días: miércoles, martes, domingo, jueves, viernes, lunes, sábado. No parece haber un patrón de mayor cantidad de lluvia durante los fines de semana (aunque la media más alta corresponde al sábado). Además, observe la gráfica de Excel de las siete medias, en donde la media del lunes se graficó primero. La gráfica de Excel no apoya la aseveración de mayor cantidad de lluvia durante los fines de semana (aunque podría argumentarse que llueve más los sábados).

Excel



- *Variación*: Las siete desviaciones estándar varían de 0.135 pulgadas a 0.290 pulgadas, pero estos valores no son muy diferentes. No parece haber algo infrecuente en las cantidades de variación.
- Los *mínimos*, *primeros cuartiles* y *medianas* son todos iguales a 0.00 para cada uno de los siete días. Lo anterior se explica por el hecho de que por cada día de la semana hay muchos días en los que no llueve. La abundancia de ceros también se observa en las gráficas de cuadro y en los histogramas, los cuales muestran que los datos tienen distribuciones cargadas hacia el extremo de los mínimos (sesgo derecho).
- *Datos distantes*: No aparecen datos distantes o valores inusuales. En el extremo de los mínimos hay muchas cantidades de lluvia iguales a cero. En el extremo de los máximos, la lista en que se ordenan las 365 cantidades de lluvia termina con los valores máximos de 0.92, 0.96, 1.28, 1.41 y 1.48.
- *Distribuciones*: Las distribuciones de las cantidades de lluvia están sesgadas hacia la derecha. No son normales, como esperaríamos. Si el uso de un método particular de estadística requiere poblaciones distribuidas normalmente (en forma de campana), este requisito no se satisface en las cantidades de lluvia.

Ahora comprendemos en gran medida la naturaleza de las cantidades de lluvia que caen en Boston durante distintos días de la semana. Con base en nuestra exploración, concluimos que en Boston no cae más lluvia durante los fines de semana que los demás días (aunque podríamos argumentar que llueve más los sábados).

Pensamiento crítico

Si nos armamos con una lista de herramientas para investigar el centro, la variación, la distribución, los datos distantes y las características de los datos a través del tiempo, tendríamos la tentación de desarrollar un procedimiento descuidado, por lo que el pensamiento crítico es sumamente importante. Además de utilizar las herramientas que se presentan en este capítulo, deberemos considerar cualesquiera otros factores que puedan ser cruciales para las conclusiones que elaboremos. En tal caso, plantearíamos preguntas como las siguientes: ¿Es posible que la muestra sea representativa de la población o está sesgada de alguna manera? ¿Cuál es la fuente de los datos? ¿Sería posible que la fuente fuera alguien con intereses que puedan afectar la calidad de los datos? Suponga, por ejemplo, que deseamos estimar el ingreso medio de estudiantes universitarios. También, suponga que enviamos por correo cuestionarios a 500 estudiantes y que recibimos 20 respuestas. Podríamos calcular la media y la desviación estándar, así como construir gráficas, identificar datos distantes, etcétera, pero los resultados serán lo que los estadísticos llaman desperdicios. La muestra es de respuesta voluntaria, por lo que no tiene posibilidades de ser representativa de la población de todos los estudiantes universitarios. Además de las herramientas estadísticas específicas presentadas en este capítulo, ¡también debemos *pensar*!

Utilizando la tecnología

Esta sección introdujo los datos distantes, los resúmenes de los cinco números y las gráficas de cuadro. Para encontrar datos distantes, se acomodan los datos en orden de menor a mayor; después, se examinan los valores máximo y mínimo para determinar si están muy lejos de los otros valores muestrales. El STATDISK, Minitab, Excel y la calculadora TI-83 Plus proporcionan valores de cuartiles, de modo que es fácil elaborar el resumen de los cinco números. El STATDISK, Minitab, Excel y la calculadora TI-83 Plus pueden utilizarse para crear gráficas de cuadro. Ahora describiremos los distintos procedimientos. (*Precaución*: Recuerde que los valores cuartilares calculados por medio de Minitab y la calculadora TI-83 Plus pueden diferir ligeramente de los calculados a partir de la figura 2-15, por lo que tal vez las gráficas de cuadro también difieran ligeramente).

STATDISK Elija el elemento **Data** del menú principal, y utilice el **Sample Editor** para introducir los datos; después, haga clic en **COPY**. Ahora seleccione **Data**, luego **Boxplot,** y haga clic en **PASTE** y en **Evaluate**.

Minitab Introduzca los datos en la columna C1; luego, seleccione **Graph** y **Boxplot**. Introduzca C1 en la primera celda, debajo de la columna Y; luego, haga clic en **OK**.

Excel Aunque Excel no se diseñó para generar gráficas de cuadro, éstas pueden crearse utilizando el Data Desk XL add-in, que complementa este libro. Primero introduzca los datos en la columna A. Haga clic en **DDXL** y seleccione **Charts y Plots.** Estando en la función Type, elija la opción de **Boxplot**. En el cuadro de diálogo, haga clic en el icono del lápiz e introduzca el rango de datos, como A1:A40, si usted tiene 40 valores listados en la columna A. Haga clic en **OK**. El resultado es una gráfica de cuadro modificada, tal como se describe en el ejercicio 13. También se muestran los valores del resumen de los cinco números.

TI-83 Plus Introduzca los datos muestrales en la lista L1. Ahora seleccione **STAT PLOT,** presionando la segunda tecla después de la tecla denominada Y =. Presione la tecla **ENTER**, después seleccione la opción **ON** y elija el tipo de gráfica de cuadro que se ubica a la mitad el segundo renglón. Xlist debe indicar L1 y el valor Freq tiene que ser 1. Ahora presione la tecla **ZOOM** y elija la opción 9 para **ZoomStat**. Presione la tecla **ENTER**; debe aparecer la gráfica de cuadro. Puede utilizar las teclas con flechas para moverse hacia la derecha o hacia la izquierda, de manera que le sea posible leer los valores desde la escala horizontal.

2-7 Destrezas y conceptos básicos

1. **Lotería** Remítase al conjunto de datos 26 y utilice sólo los 40 dígitos en la primera columna de los resultados *Win 4* de la lotería del estado de Nueva York (9, 7, 0, etcétera). Encuentre el resumen de los cinco números y construya una gráfica de cuadro. ¿Qué características de la gráfica de cuadro sugieren que los dígitos fueron seleccionados con un procedimiento aleatorio y justo?

2. Presupuestos de películas Remítase al conjunto de datos 21 del Apéndice B, con los montos de presupuesto de las 15 películas con clasificación R. Elabore el resumen de los cinco números y construya una gráfica de cuadro. Determine si los valores muestrales son representativos de las películas realizadas este año.

3. Calorías de cereales Remítase al conjunto de datos 16 del Apéndice B de los 16 valores de las calorías por gramo de cereales. Elabore el resumen de los cinco números y construya una gráfica de cuadro. Determine si los valores muestrales serían representativos de los cereales consumidos por la población en general.

4. Nicotina en cigarrillos Remítase al conjunto de datos 5 de las 29 cantidades de nicotina (en miligramos por cigarrillo). Elabore el resumen de los cinco números y construya una gráfica de cuadro. ¿Podrían los valores muestrales ser representativos de los cigarrillos fumados por un consumidor individual?

5. M&M rojos Remítase al conjunto de datos 9 de los 21 pesos (en gramos) de los dulces M&M rojos. Elabore el resumen de los cinco números y construya una gráfica de cuadro. ¿Podrían los valores muestrales ser representativos de los dulces M&M de todos los colores?

T **6.** Longitudes de osos Remítase al conjunto de datos 9 de las longitudes (en pulgadas) de los 54 osos que anestesiamos y medimos. Elabore el resumen de los cinco números y construya una gráfica de cuadro. ¿Podría la distribución de longitudes ser simétrica? ¿O está sesgada?

T **7.** Alcohol en películas infantiles Remítase al conjunto de datos 7 de las 50 duraciones (en segundos) de escenas que presentan consumo de alcohol en películas infantiles de dibujos animados. Elabore el resumen de los cinco números y construya una gráfica de cuadro. Con base en la gráfica de cuadro, ¿la distribución parece simétrica o está sesgada?

T **8.** Temperaturas corporales Remítase al conjunto de datos 4 del Apéndice B de las 106 temperaturas corporales a las 12 a. m. del día 2. Elabore el resumen de los cinco números y construya una gráfica de cuadro; después, determine si los valores muestrales apoyan la creencia común de que la temperatura corporal media es de 98.6°F.

En los ejercicios 9 a 12 elabore los resúmenes de los cinco números, construya gráficas de cuadro y compare los conjuntos de datos.

9. Premios Óscar En el artículo "Ages of Oscar Winning Best Actors and Actresses" (revista *Mathematics Teacher*), de Richard Brown y Gretchen Davis, los autores comparan las edades de actores y actrices en el momento de ganar el Óscar. En la siguiente tabla se presentan los resultados de los ganadores de ambas categorías. Utilice gráficas de cuadro para comparar los dos conjuntos de datos.

Actores:	32	37	36	32	51	53	33	61	35	45	55
	39	76	37	42	40	32	60	38	56	48	48
	40	43	62	43	42	44	41	56	39	46	31
	47	45	60	46	40	36					
Actrices:	50	44	35	80	26	28	41	21	61	38	49
	33	74	30	33	41	31	35	41	42	37	26
	34	34	35	26	61	60	34	24	30	37	31
	27	39	34	26	25	33					

T **10.** Coca Cola regular/Coca Cola dietética Remítase al conjunto de datos 17 del Apéndice B; utilice los pesos de la Coca Cola regular y los pesos de la Coca Cola dietética. ¿Parece haber una diferencia significativa? Si es así, ¿qué explicación le encuentra?

11. Niveles de cotinina Remítase a la tabla 2-1 ubicada en el problema del capítulo. Ya calculamos que el resumen de los cinco números para los niveles de cotinina de fumadores es 0, 86.5, 170, 251.5 y 491. Elabore los resúmenes de los cinco números para los otros dos grupos; después, construya las tres gráficas de cuadro utilizando la misma escala. ¿Existe alguna diferencia aparente?

12. Clancy, Rowling, Tolstoi Remítase al conjunto de datos 14 del Apéndice B y utilice las puntuaciones de la escala de facilidad de lectura de Flesch para las páginas muestra de las obras *El oso y el dragón*, de Tom Clancy; *Harry Potter y la piedra filosofal*, de J. K. Rowling, y *La guerra y la paz*, de León Tolstoi. (Las puntuaciones más altas indican una lectura más fácil). ¿Parece haber alguna diferencia en la facilidad de lectura? ¿Son los resultados consistentes con sus expectativas?

2-7 Más allá de lo básico

13. Las gráficas de cuadro introducidas en esta sección suelen denominarse gráficas de cuadro *de esqueleto* (o *regulares*). Las **gráficas de cuadro modificadas** se construyen de la siguiente forma:

a. Calcule el *RIC*, que denota el rango intercuartilar, definido por $RIC = Q_3 - Q_1$.

b. Dibuje el cuadro con la mediana y los cuartiles como siempre; pero, cuando trace las líneas a la derecha e izquierda del cuadro, dibújelas sólo hasta los puntos que corresponden a los valores máximo y mínimo, que están dentro de 1.5 *RIC* del cuadro.

c. Los **datos ligeramente distantes**, que se grafican como puntos *sólidos*, son valores que están por debajo de Q_1 o por arriba de Q_3, por una cantidad mayor que 1.5 *RIC*, pero no mayor que 3 *RIC*. Es decir, los datos ligeramente distantes son valores *x*, tales que

$$Q_1 - 3\,RIC \leq x < Q_1 - 1.5\,RIC$$

o

$$Q_3 + 1.5\,RIC < x \leq Q_3 + 3\,RIC$$

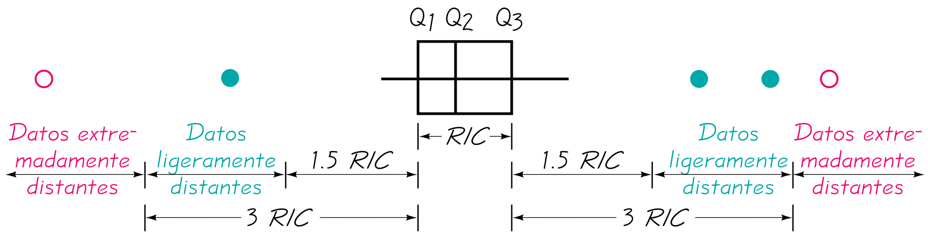
d. Los **datos extremadamente distantes**, que se grafican como pequeños círculos *vacíos*, son valores que están por debajo de Q_1 por más de 3 *RIC* o por encima de Q_3 por más de 3 *RIC*. Es decir, los datos extremadamente distantes son valores *x*, tales que

$$x < Q_1 - 3\,RIC$$

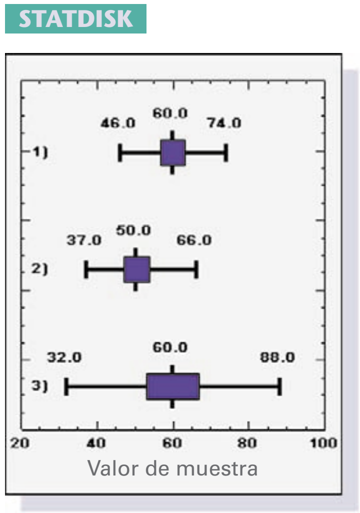
o

$$x > Q_3 + 3\,RIC$$

La figura anexa es un ejemplo de una gráfica de cuadro modificada. Remítase a los niveles de cotinina de fumadores de la tabla 2-1, incluida en el problema del capítulo. Ya sabemos que este conjunto de datos tiene un resumen de los cinco números de 0, 86.5, 170, 251.5 y 491. Identifique el valor de *RIC*, identifique los rangos de valores utilizados para localizar sus datos ligeramente distantes y extremadamente distantes; después, identifique cualesquiera datos ligeramente distantes y extremadamente distantes reales.



14. Remítase a la representación visual de STATDISK, de las tres gráficas de cuadro que representan la medida de longevidad (en meses) de muestras de tres distintas baterías para automóvil. Si usted es el encargado de una flotilla de automóviles y debe seleccionar una de las tres marcas, ¿cuál gráfica de cuadro representa la marca que debería elegir? ¿Por qué?



Repaso

En este capítulo consideramos métodos para describir, explorar y comparar conjuntos de datos. Cuando se investigan conjuntos de datos, las siguientes características son, por lo general, muy importantes:

1. *Centro*: Un valor representativo o promedio.
2. *Variación*: Una medida de la cantidad en que varían los valores.
3. *Distribución*: La naturaleza o forma de la distribución de los datos (como normal, uniforme o sesgada).
4. *Datos distantes*: Valores muestrales que se ubican muy lejos de la mayoría de valores muestrales.
5. *Tiempo*: Características cambiantes de los datos a través del tiempo.

Después de completar este capítulo, usted será capaz de hacer lo siguiente:

- Resumir datos por medio de la construcción de una distribución de frecuencias o de una distribución de frecuencias relativas (sección 2-2).
- Representar visualmente la naturaleza de la distribución por medio de la construcción de un histograma, una gráfica de puntos, una gráfica de tallo y hojas, una gráfica circular o una gráfica de Pareto (sección 2-3).
- Calcular medidas de tendencia central como la media, la mediana, la moda y la mitad del rango (sección 2-4).
- Calcular medidas de variación como la desviación estándar, la varianza y el rango (sección 2-5).
- Comparar valores individuales utilizando puntuaciones z, cuartiles o percentiles (sección 2-6).
- Investigar y explorar la dispersión de los datos, el centro de los datos y el rango de los valores por medio de la construcción de una gráfica de cuadro (sección 2-7).

Además de crear dichas tablas, gráficas y medidas, usted será capaz de comprender e interpretar los resultados. Por ejemplo, debe entender con claridad que la desviación estándar es una medida acerca de qué tanto varían los datos, y saber utilizar la desviación estándar para distinguir entre valores frecuentes e infrecuentes.

Ejercicios de repaso

1. Edades de presidentes La senadora Hayes está considerando competir por la presidencia de Estados Unidos, pero sólo tiene 35 años de edad, que es la edad mínima requerida. Al investigar este tema, descubre las edades de presidentes anteriores en el momento de tomar el cargo; dichas edades se listan abajo. Utilice las edades y calcule *a*) la media; *b*) la mediana; *c*) la moda; *d*) la mitad del rango; *e*) el rango; *f*) la desviación estándar; *g*) la varianza; *h*) Q_1; *i*) Q_3; y *j*) P_{10}.

57	61	57	57	58	57	61	54	68	51	49	64	50	48
65	52	56	46	54	49	51	47	55	55	54	42	51	56
55	51	54	51	60	62	43	55	56	61	52	69	64	46
54													

2. **a.** John F. Kennedy tenía 43 años de edad cuando tomó posesión. Utilice los resultados del ejercicio 1 y convierta esta edad a puntuación *z*.
b. ¿Será la edad de 43 años de Kennedy “infrecuente”? ¿Por qué?
c. Aplique la regla práctica del intervalo para identificar otras edades de la lista que sean infrecuentes.
d. Aun cuando la lista de edades no incluye la de 35 años, ¿sería esa edad infrecuente? ¿Es probable que un candidato a la presidencia, de 35 años de edad, descubra que su edad sea un tema importante de campaña?

3. Distribución de frecuencias Utilice la misma lista de edades del ejercicio 1 y construya una distribución de frecuencias. Use seis clases, con 40 como el límite inferior de la primera clase, y una anchura de clase de 5.

4. Histograma Utilice la distribución de frecuencias del ejercicio 3, construya un histograma e identifique la naturaleza general de la distribución (ya sea uniforme, normal, sesgada).

5. Gráfica de cuadro Utilice las mismas edades de la lista en el problema 1, construya una gráfica de cuadro e identifique los valores que constituyen el resumen de los cinco números.

6. Regla empírica Suponga que las edades de presidentes pasados, presentes y futuros tienen una distribución normal, con una media de 54.8 años y una desviación estándar de 6.2 años.
a. ¿Qué dice la regla empírica acerca del porcentaje de edades entre 48.6 años y 61.0 años (o dentro de una desviación estándar de la media)?
b. ¿Qué dice la regla empírica acerca del porcentaje de edades entre 42.4 y 67.2 años?

7. Comparación de puntuaciones Un psicólogo industrial de la Citation Corporation crea dos pruebas diferentes para medir la satisfacción laboral. ¿Cuál puntuación es mejor: una de 72 en la prueba de administración, la cual tiene una media de 80 y una desviación estándar de 12, o una de 19 en la prueba de producción de empleados, con una media de 20 y una desviación estándar de 5? Explique.

8. **a.** Estime la media de la edad de los automóviles que conducen los estudiantes de su universidad.
b. Utilice la regla práctica del intervalo para hacer un estimado de la desviación estándar de la edad de los automóviles que conducen los estudiantes de su universidad.

9. Transformación de datos Una profesora de estadística encontró que el tiempo que emplean los estudiantes para resolver su examen final tiene una media de 135 minutos y una desviación estándar de 15 minutos. Planea añadir una nueva pregunta, la cual requerirá cinco minutos adicionales de cada estudiante.
a. ¿Cuál es la media después de incluir la nueva pregunta?
b. ¿Cuál es la desviación estándar después de incluir la nueva pregunta?
c. ¿Cuál es la varianza después de incluir la nueva pregunta?

10. Quejas de pasajeros aéreos En un año reciente hubo 23,000 quejas por parte de pasajeros aéreos. Las categorías y frecuencias de las quejas, proporcionadas por el Departamento de Transporte de Estados Unidos, son las siguientes: atención al cliente (4370); problemas con el vuelo (9200); reservaciones, boletaje y abordaje (1610); equipaje (3450); obtención de reembolsos (1150), y otras razones (3220). Construya una gráfica de Pareto que resuma los datos.

Ejercicios de repaso acumulativos

1. Errores del reloj de pulso Como parte de un proyecto para la clase de estadística, un estudiante reúne datos sobre la precisión de relojes de pulso; obtiene los siguientes errores del tiempo (en segundos). (Los valores positivos representan relojes que se adelantan; los valores negativos representan relojes que se atrasan).

140 −125 105 −241 −85 41 186 −151 325 80 27 20 20 30 −65

a. Calcule la media, la mediana, la moda y la mitad del rango.
b. Calcule la desviación estándar, la varianza y el rango.
c. ¿Provendrán los tiempos dados de una población discreta o de una continua?
d. ¿Cuál es el nivel de medición de estos valores (nominal, ordinario, de intervalo, de razón)?

2. a. Un conjunto de datos tiene un nivel nominal de medición y usted desea obtener un dato representativo. ¿Cuál de los siguientes es el más apropiado: la media, la mediana, la moda o la mitad del rango? ¿Por qué?
b. Se obtiene una muestra al llamar por teléfono a las primeras 250 personas listadas en el directorio telefónico local. ¿Qué tipo de muestreo se utilizó (aleatorio, estratificado, sistemático, de racimo, de conveniencia)?
c. Se lleva a cabo una encuesta de salida, consistente en encuestar a cada persona que sale de la casilla electoral en 50 distritos seleccionados aleatoriamente. ¿Qué tipo de muestreo se utilizó (aleatorio, estratificado, sistemático, de racimo, de conveniencia)?
d. Un fabricante recarga cartuchos para impresoras de computadora. Un gerente descubre que la cantidad de tinta vertida en el recipiente no es muy consistente, de tal forma que algunos cartuchos duran más de lo esperado, en tanto que otros se agotan demasiado pronto. Él desea mejorar la calidad haciendo que la cantidad de tinta en los cartuchos tenga más consistencia. Cuando se analizan las cantidades de tinta, ¿cuál de los siguientes estadísticos es más importante: la media, la mediana, la moda, la mitad del rango, la desviación estándar, el primer cuartil, el tercer cuartil? ¿Debería elevarse, disminuirse o dejar sin cambio el valor de dicho estadístico?

3. Consumo de energía Cada año, el Departamento de Energía de Estados Unidos publica un *Annual Energy Review*, que incluye el consumo de energía *per capita* (en millones de BTU) de cada uno de sus 50 estados. Si se calcula la media de estos 50 valores, ¿el resultado será la media de consumo de energía *per capita* de la población de los 50 estados combinados? Si no es así, explique cómo calcularía la media del consumo de energía *per capita* de los 50 estados combinados.

Actividades de cooperación en equipo

1. Actividad fuera de clase ¿Influyen las cifras de anclaje en los estimados? En el artículo “Weighing Anchors”, de la revista *Omni*, John Rubin observó que, cuando la gente estima un valor, su estimación suele estar “anclada a” (o influida por) un número anterior, aun cuando ese número no tenga ninguna relación con la cantidad que se estima. Para demostrar esto, pidió a varias personas que le dieran un estimado rápido del valor de $8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$. La respuesta promedio fue 2250, pero cuando se invirtió el orden de los números el promedio fue 512. Rubin explicó que cuando iniciamos un cálculo con números grandes (como en $8 \times 7 \times 6$), nuestros estimados tienden a ser grandes. Señaló que tanto 2250 como 512 son mucho menores que el producto correcto, 40,320. El artículo sugiere que números irrelevantes pueden incluirse en los avalúos de bienes raíces, así como en las estimaciones del valor de un automóvil y en las de la posibilidad de una guerra nuclear.

   Realice un experimento para probar dicha teoría. Seleccione a algunos sujetos y pídales que estimen rápidamente el valor de

$$8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$$

   Después, seleccione otros sujetos y pídales que estimen rápidamente el valor de

$$1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8$$

   Registre los estimados junto con el orden utilizado. Diseñe con cuidado el experimento, de modo que las condiciones sean uniformes y que los dos grupos muestrales se seleccionen de forma que se minimice cualquier sesgo. No describa la teoría a los sujetos hasta después de proporcionar sus estimaciones. Compare los dos conjuntos de resultados muestrales, a través del uso de los métodos de este capítulo. Prepare un reporte impreso, que incluya los datos reunidos, los métodos detallados utilizados, el método de análisis, las gráficas y/o los estadísticos relevantes, así como las conclusiones. Incluya una crítica dando razones por las que los resultados podrían ser incorrectos y describa formas para mejorar el experimento.

   Una variante del experimento anterior consiste en encuestar personas acerca de su conocimiento sobre la población de Kenia. Primero pregúntele a la mitad de los sujetos si piensa que la población es mayor o menor que cinco millones; después, pídales que estimen la población dando un número real. Pregunte a la otra mitad de los sujetos si cree que la población es mayor o menor que 80 millones; después, pídales que estimen la población. (La población de Kenia es de 28 millones). Compare los dos conjuntos de resultados e identifique el efecto “de anclaje” en la cifra inicial que dieron los sujetos encuestados.

2. Actividad fuera de clase Cada equipo, formado por tres o cuatro estudiantes, debe reunir un conjunto original de datos que estén a un nivel de medición de intervalo o de razón. Proporcione lo siguiente: *a*) una lista de valores muestrales, *b*) resultados de computadora impresos con estadísticos descriptivos y gráficas, y *c*) una descripción por escrito de la naturaleza de los datos, el método de recolección y las características importantes.

3. Actividad en clase A continuación se indican las edades que un grupo de motociclistas tenía cuando se hirió fatalmente en accidentes de tránsito (según datos del Departamento del Transporte de Estados Unidos). Si su objetivo es dramatizar el peligro que constituyen las motocicletas para la gente joven, ¿cuál de los siguientes sería el más efectivo: el histograma, la gráfica de Pareto, la gráfica circular, la gráfica de puntos, la media, la mediana? Construya la gráfica y encuentre el estadístico que cumple mejor el objetivo. ¿Es correcto distorsionar los datos de manera deliberada si el objetivo es salvar la vida de los motociclistas?

17	38	27	14	18	34	16
42	28	24	40	20	23	31
37	21	30	25	17	28	33
25	23	19	51	18	29	

4. Actividad fuera de clase Pida a cada equipo, que se formó con tres o cuatro estudiantes, que seleccione uno de los siguientes reactivos y que construya una gráfica que sirva para poner énfasis en la pregunta:
   a. ¿Habrá una diferencia entre los valores del índice de masa corporal (IMC) de los hombres y el de las mujeres? (Véase el conjunto de datos 1 del Apéndice B).
   b. ¿Existirá alguna relación entre la estatura de los hijos (o hijas) y la estatura de sus padres (o madres)? (Véase el conjunto de datos 2 del Apéndice B).
   c. ¿Parecen los dígitos generados por la lotería *Win 4* del estado de Nueva York haber sido seleccionados aleatoriamente? ¿O estarán sesgados? (Véase el conjunto de datos 26 del Apéndice B).

Proyecto tecnológico

Cuando se manejan conjuntos grandes de datos, la introducción manual de éstos suele ser tediosa y requerir de mucho tiempo. Hay mejores actividades que hacer con su tiempo, como aprender los principios aerodinámicos de un *frisbee*. Remítase al conjunto de datos 30 del Apéndice B, que incluye las distancias de los jonrones “”de tres jugadores de beisbol excepcionales: Barry Bonds (temporada 2001), Mark McGwire (temporada 1998) y Sammy Sosa (temporada 1998). En lugar de introducir manualmente las 209 distancias de los tres conjuntos de datos, utilice la calculadora TI-83 Plus, el STATDISK, el Minitab o Excel para cargar el conjunto de datos, los cuales están disponibles en el CD incluido en este libro. Proceda a generar histogramas, obtendrá estadísticos apropiados que le permitirán comparar los tres conjuntos de datos. ¿Hay algunas diferencias significativas? ¿Existen algunos datos distantes? ¿Parece que los jugadores que golpean más lejos hacen más jonrones””? ¿Por qué? Analice los últimos dígitos de las distancias y determine si los valores parecen ser estimaciones o mediciones. Escriba un reporte breve que contenga sus conclusiones y gráficas de apoyo.

Conjunto de datos 30: Distancias de jonrones

Las distancias de los jonrones de Mark McGwire (1998), Sammy Sosa (1998) y Barry Bonds (2001) están en pies.

Nombre de los archivos de STATDISK y de texto: MCGWR, SOSA, BONDS.
Minitab: el nombre de la hoja de cálculo es HOMERUNS.MTW.
Excel: el nombre del libro es HOMERUNS.XLS.
TI-83 Plus: el nombre es HOMERUNS, y los nombres de los archivos son los mismos de STATDISK y de los archivos de texto.

McGwire

360	370	370	430	420	340	460	410	440	410
380	360	350	527	380	550	478	420	390	420
425	370	480	390	430	388	423	410	360	410
450	350	450	430	461	430	470	440	400	390
510	430	450	452	420	380	470	398	409	385
369	460	390	510	500	450	470	430	458	380
430	341	385	410	420	380	400	440	377	370

Sosa

371	350	430	420	430	434	370	420	440	410
420	460	400	430	410	370	370	410	380	340
350	420	410	415	430	380	380	366	500	380
390	400	364	430	450	440	365	420	350	420
400	380	380	400	370	420	360	368	430	433
388	440	414	482	364	370	400	405	433	390
480	480	434	344	410	420				

Bonds

420	417	440	410	390	417	420	410	380	430
370	420	400	360	410	420	391	416	440	410
415	436	430	410	400	390	420	410	420	410
410	450	320	430	380	375	375	347	380	429
320	360	375	370	440	400	405	430	350	396
410	380	430	415	380	375	400	435	420	420
488	361	394	410	411	365	360	440	435	454
442	404	385							

de los DATOS *a la* DECISIÓN

Pensamiento crítico

Las muertes por choques de automóviles son devastadoras para las familias de las víctimas y con frecuencia implican procesos legales y pagos de seguro costosos. A continuación se presentan las edades de 100 conductores que murieron en choques de automóviles, seleccionados aleatoriamente. También se incluye una distribución de frecuencias, por edades, de conductores con licencia.

Edad (en años) de conductores muertos en choques de automóviles

37	76	18	81	28	29	18	18	27	20
18	17	70	87	45	32	88	20	18	28
17	51	24	37	24	21	18	18	17	40
25	16	45	31	74	38	16	30	17	34
34	27	87	24	45	24	44	73	18	44
16	16	73	17	16	51	24	16	31	38
86	19	52	35	18	18	69	17	28	38
69	65	57	45	23	18	56	16	20	22
77	18	73	26	58	24	21	21	29	51
17	30	16	17	36	42	18	76	53	27

Edad	Conductores con licencia (millones)
16–19	9.2
20–29	33.6
30–39	40.8
40–49	37.0
50–59	24.2
60–69	17.5
70–79	12.7
80–89	4.3

Análisis

Convierta la distribución de frecuencias a una distribución de frecuencias relativas; después, elabore una distribución de frecuencias relativas con las edades de los conductores que murieron en choques de automóviles. Compare las dos distribuciones de frecuencias relativas. ¿Cuáles categorías de edad parecen tener proporciones sustancialmente mayores de muertes que las proporciones de los conductores con licencia? Si usted fuese el responsable de establecer las tasas de seguros de automóviles, ¿a qué categorías de edad les asignaría las tasas más altas? Construya una gráfica que permita identificar las categorías de edad más propensas a accidentes automovilísticos fatales.

PROYECTO DE INTERNET

Datos en Internet

Internet posee una enorme cantidad de información, de la cual, gran parte, se presenta en forma de datos brutos que pueden estudiarse y resumirse con el uso de los estadísticos presentados en este capítulo. Por ejemplo, encontramos la siguiente información con tan sólo unos cuantos clics:

- El valor de la acción de Walt Disney Corporation tiende a subir durante los meses de invierno, pero varía el resto del año. En el 2001, los precios máximo y mínimo de la acción variaron más de 17 puntos.
- Ichiro Suzuki, jardinero lateral del equipo de beisbol de los Marineros de Seattle, tuvo un porcentaje de "hits" de 0.457 durante la temporada 2001.
- Las poblaciones de California, Nueva York y Texas representan más del 25% del total de la de Estados Unidos.

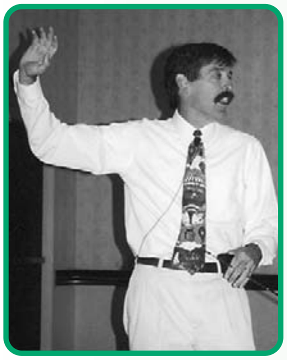
El proyecto de Internet para este capítulo, que se encuentra en el sitio de Internet de *Estadística elemental*, lo conducirá a conjuntos de datos en las áreas de deportes, finanzas y clima. Una vez que haya armado un conjunto de datos, aplicará los métodos de este capítulo para resumir y clasificar los datos.

El sitio Web para este capítulo se encuentra en

www.pearsoneducacion.net/triola

La estadística @ en el trabajo

"El periodista debe ser capaz de tener una visión crítica hacia la investigación, y comprender el verdadero contexto y el significado del trabajo".



Mark Fenton

Editor de la Revista Large Walking Magazine

Mark Fenton también es defensor de la marcha y campeón de este deporte. Perteneció al equipo nacional de marcha de Estados Unidos en cinco ocasiones, y ha representado a este país en numerosas competencias internacionales. Estudió biomecánica y fisiología del deporte en el Olympic Training Center's Sports Science Laboratory, ubicado en Colorado Springs, Colorado.

¿Cuáles conceptos de estadística utiliza?

Debo estar familiarizado con todas las herramientas comunes que se utilizan en los análisis estadísticos de las ciencias del deporte y de la investigación en salud pública; desde las medias y las desviaciones estándar, hasta la significancia o diferencias estadísticas, los intervalos de confianza y el análisis de varianza, entre otros.

¿De qué manera utiliza la estadística en su trabajo?

Suelo leer revistas de investigaciones médicas (*Medicine and Science in Sports and Exercise* y el *Journal of the American Medical Association* son los más relevantes) sobre fisiología del deporte, salud pública e investigación epidemiológica. Debo mantenerme informado sobre la manera en que se controlan y analizan los estudios, así como comprender la potencia o el valor relativo del resultado de un estudio. Esta información la utilizo en los artículos y conferencias que elaboro sobre los hallazgos de dicho trabajo.

Por favor, describa un ejemplo específico que ilustre la forma en que el uso de la estadística haya logrado mejorar un producto o servicio.

Por lo general, leo artículos de investigación y debo estar atento a los vacíos de la evidencia, tales como tamaños de muestra demasiado pequeños; significancia estadística con diferencias absolutas pequeñas entre grupos; desviaciones estándar extremadamente grandes; todo lo cual puede oscurecer la relevancia de un hallazgo para una persona promedio.

Un ejemplo simple es el problema que existe al afirmar que el pulso máximo **promedio** de una mujer de 36 años es de 190 latidos por minuto (226 menos la edad). Esto es útil para calcular el pulso que se busca, hasta que uno encuentra que la desviación es demasiado alta y la cifra resultante puede alejarse tanto como 10 latidos por minuto para casi 1/3 de la población, un error suficientemente grande para causar problemas al hacer ejercicio. El estudio que señaló esto alteró la forma en que recomiendo la intensidad del ejercicio.

¿Se estará incrementando o disminuyendo el uso de la probabilidad y la estadística, o permanece estable?

Debido a mi trabajo cada vez más intenso en salud pública y al incremento en la complejidad de las herramientas estadísticas que se utilizan en los análisis de poblaciones grandes, mis conocimientos en este campo **deben** seguir acrecentándose si quiero mantenerme al día.

¿Recomienda el uso de la estadística a los estudiantes universitarios de hoy?

Recomiendo al menos un curso de estadística, ya que es una herramienta útil para evaluar la información con que se nos bombardea cada día, mucha de la cual carece de contexto. Cualquier persona que se interesa en la ciencia del periodismo debe tomar al menos un curso introductorio extenso de estadística. El periodista debe ser capaz de tener una visión crítica hacia la investigación, y comprender el verdadero contexto y el significado del trabajo.

3

Probabilidad

PROBLEMA 3 DEL CAPÍTULO

Falsos positivos y falsos negativos

En diferentes etapas de nuestra vida, todos nos sometemos a una variedad de exámenes médicos. Algunos exámenes médicos son tan simples, como aquellos donde se usa un termómetro para establecer si la temperatura corporal es muy alta o muy baja, o como un esfigmomanómetro con el que se determina si la presión sanguínea es muy alta o muy baja. Otros exámenes clínicos incluyenn el análisis de muestras de sangre para identificar la presencia de alguna enfermedad. En este problema del capítulo consideramos los resultados obtenidos en un estudio clínico consistente en una prueba de embarazo. Para una mujer es importante saber si está embarazada con la finalidad de interrumpir prácticas que serían potencialmente dañinas para el bebé, como las actividades físicas, la medicación, la exposición a tóxicos en el trabajo, el tabaquismo o el consumo de alcohol. Las pruebas de embarazo, como casi todas las pruebas médicas, arrojan resultados que distan de ser 100% precisos. Los resultados mostrados en la tabla 3-1 se obtuvieron con la prueba de embarazo de Abbot, a partir de muestras de sangre (según datos de "Specificity and Detection Limit of Ten Pregnancy Tests", de Tiitinen y Stenman, *Scandinavian Journal of Clinical Laboratory Investigation*, vol. 53, suplemento 216). Existen factores, como el avance del embarazo, que afectan la precisión de dichos exámenes. Las pruebas de embarazo son, por lo regular, más confiables cuando se aplican al menos dos semanas después de la concepción. Otras pruebas ofrecen resultados más confiables que los de la tabla 3-1. Por ejemplo, la de Abbot Testpack Plus es una prueba de orina con una tasa de falso positivo de 0.2% y una tasa de falso negativo de 0.6%. Los términos *falso positivo* y *falso negativo* se incluyen entre los siguientes términos, que se usan comúnmente en las pruebas médicas o en procedimientos de vigilancia:

- *Falso positivo*: La prueba indica incorrectamente embarazo cuando la mujer no está embarazada.
- *Falso negativo*: La prueba indica incorrectamente que la mujer no está embarazada cuando en realidad lo está.
- *Verdadero positivo*: La prueba indica correctamente que la mujer está embarazada cuando en realidad lo está.
- *Verdadero negativo*: La prueba indica correctamente que la mujer no está embarazada cuando no lo está.
- *Sensibilidad de la prueba*: La probabilidad de un verdadero positivo.
- *Especificidad de la prueba*: La probabilidad de un verdadero negativo.

Con base en los resultados de la tabla 3-1, ¿cuál es la probabilidad de que una mujer esté embarazada si la prueba indica un resultado negativo? ¿Cuál es la probabilidad de un falso positivo? En este capítulo resolveremos estas preguntas.

Tabla 3-1 Resultados de prueba de embarazo

	Resultado de prueba positivo (indicó embarazo)	Resultado de prueba negativo (no indicó embarazo)
La mujer está embarazada	80	5
La mujer no está embarazada	3	11

3-1 Panorama general

La probabilidad es la base sobre la que se construyen los métodos importantes de la estadística inferencial. Como un sencillo ejemplo, suponga que usted hubiera ganado el premio mayor de la lotería nacional cinco veces seguidas. Habría acusaciones de que usted hizo trampa de alguna forma. Las personas saben que aun cuando existe la posibilidad de que alguien gane cinco veces consecutivas, por pura suerte, la posibilidad es tan increíblemente baja, que rechazarían la suerte como una explicación razonable. Ésta es precisamente la forma de pensar de los estadísticos: las personas rechazan las explicaciones basadas en probabilidades muy bajas. Los estadísticos usan la *regla del suceso infrecuente*.

Regla del suceso infrecuente para estadística inferencial

Si, bajo un supuesto dado (como un juego de lotería justo), la probabilidad de un suceso particular observado (como ganar cinco veces consecutivas) es extremadamente pequeña, concluimos que el supuesto probablemente es incorrecto.

El objetivo principal de este capítulo es desarrollar una comprensión válida de los valores de probabilidad, la cual se usará en los capítulos siguientes. Un objetivo secundario es desarrollar las habilidades básicas necesarias para determinar los valores de probabilidad en una variedad de circunstancias importantes.

3-2 Fundamentos

Al considerar la probabilidad, tratamos con procedimientos (como tirar un dado, contestar una pregunta de opción múltiple en un examen o aplicar una prueba de embarazo) que producen resultados.

Definiciones

Suceso: cualquier conjunto de resultados o consecuencias de un procedimiento.

Un **suceso simple** es un resultado o un suceso que ya no puede desglosarse en componentes más simples.

El **espacio muestral** de un procedimiento se compone de todos los sucesos *simples* posibles. Es decir, el espacio muestral se forma con todos los resultados que ya no es posible desglosar más.

EJEMPLOS

Procedimiento	Ejemplo de suceso	Espacio muestral
Tirar un dado	5 (suceso simple)	{1, 2, 3, 4, 5, 6}
Tirar dos dados	7 (suceso no simple)	{1-1, 1-2, . . . , 6-6}

Cuando se tira un dado, el resultado 5 es un *suceso simple* porque no es posible desglosarlo en otros. Cuando se tiran dos dados, el resultado 7 *no es un suceso simple*, porque esto puede todavía desglosarse en eventos más simples, tales como 3-4 o 6-1. Cuando tiramos dos dados, el resultado de 3-4 se considera un suceso simple, porque no es posible desglosarlo más. Tal vez pensaríamos de forma incorrecta que 3-4 se desglosaría en los resultados individuales de 3 y 4, pero, cuando se tiran dos dados, 3 y 4 no son resultados individuales. Cuando se tiran dos dados, existen exactamente 36 resultados que son sucesos simples: 1-1, 1-2, . . . , 6-6.

Hay diferentes formas para definir la probabilidad de un suceso; expondremos tres enfoques. Para iniciar, presentamos una lista de algunas notaciones básicas.

Notación de probabilidades

P denota una probabilidad.
A, B y C denotan sucesos específicos.
$P(A)$ denota la probabilidad de que ocurra el suceso A.

Regla 1: Aproximación de la probabilidad por frecuencias relativas

Realice (u observe) un procedimiento un gran número de veces y cuente las ocasiones que el suceso A ocurre en realidad. Con base en estos resultados reales, $P(A)$ se estima de la siguiente forma:

$$P(A) = \frac{\text{número de veces que ocurre } A}{\text{número de veces que se repitió el ensayo}}$$

Regla 2: Método clásico de la probabilidad (requiere resultados igualmente probables)

Suponga que un procedimiento dado tiene n sucesos simples distintos, cada uno de los cuales tiene la misma posibilidad de ocurrir. Si el suceso A puede ocurrir en s de estas n formas, entonces

$$P(A) = \frac{\text{número de formas en que puede ocurrir } A}{\text{número de sucesos simples diferentes}} = \frac{s}{n}$$

Regla 3: Probabilidades subjetivas

$P(A)$, la probabilidad del suceso A, se obtiene simplemente suponiendo o estimando su valor con base en el conocimiento de las circunstancias relevantes.

LA ESTADÍSTICA EN LAS NOTICIAS

Asteroides asesinos

La probabilidad de que nuestra civilización sea destruida por un asteroide que choque con nuestro planeta tiene una importancia evidente para casi todos nosotros. En junio de 2002 un artículo del *New York Times* reportó que un asteroide "suficientemente grande como para arrasar una gran ciudad se acercó hasta 75,000 millas (unos 120,000 kilómetros) de la Tierra... pero no fue detectado hasta pasados tres días". Cuando se intenta determinar esa probabilidad, el método de las frecuencias relativas no se aplica, porque es imposible realizar ensayos y no hay datos históricos de una destrucción de ese tipo. El método clásico no se aplica porque los resultados posibles no tienen la misma probabilidad de ocurrir. Sólo puede aplicarse el método de la probabilidad subjetiva.

Con base en observaciones más actuales, los astrónomos estiman que hay 700,000 asteroides lo suficientemente grandes y cercanos como para destruirnos. Al usar este número y el conocimiento de las órbitas de los asteroides, los astrónomos desarrollaron la probabilidad subjetiva de que a nuestra civilización la destruya una colisión con un asteroide en algún momento en los próximos 100 años: la probabilidad es aproximadamente de 1/5000.

El cumpleaños más común: el 5 de octubre

Un sitio Web indicó que "una investigación reciente en bases de datos, realizada por Anybirthday. com, sugiere que el 5 de octubre es la fecha de nacimiento más popular en Estados Unidos". Se notó que una concepción en la noche de Año Nuevo podría resultar probablemente en un nacimiento el 5 de octubre. La fecha de nacimiento menos común se identificó como el 22 de mayo. Al parecer, el 18 de agosto no tiene el mismo encanto que la noche de Año Nuevo.

(*a*) (*b*) (*c*)

FIGURA 3-1 Tres métodos para calcular la probabilidad

a) Método de las frecuencias relativas (regla 1). Cuando se trata de determinar: *P* (tachuela cae con la punta hacia arriba), debemos repetir muchas veces el procedimiento de lanzar la tachuela y después calcular el cociente del número de veces que la tachuela cae con la punta hacia arriba entre el número de lanzamientos.

b) Método clásico (regla 2). Cuando se trata de determinar *P*(2) con un dado balanceado, cada una de las seis caras tiene la misma probabilidad de ocurrir.

$$P(2) = \frac{\text{número de formas en que 2 puede ocurrir}}{\text{número total de sucesos simples}}$$

$$= \frac{1}{6}$$

c) Probabilidad subjetiva (regla 3). Cuando se trata de estimar la probabilidad de que mañana llueva, los meteorólogos usan su conocimiento experto de las condiciones del tiempo para desarrollar un estimado de la probabilidad.

Es muy importante notar que *el método clásico (regla 2) requiere resultados igualmente probables*. Si los resultados no son igualmente probables, debemos usar el estimado de frecuencias relativas o confiar en nuestro conocimiento de las circunstancias para hacer una *conjetura entrenada*. La figura 3-1 ilustra los tres métodos.

Al calcular probabilidades con el método de frecuencias relativas (regla 1), obtenemos un *estimado* en lugar de un valor exacto. Conforme el número total de observaciones se incrementa, los estimados correspondientes tienden a acercarse a la probabilidad real. Tal propiedad se enuncia en forma de teorema, al que se conoce comúnmente como la *ley de los grandes números*.

Ley de los grandes números

Conforme un procedimiento se repite una y otra vez, la probabilidad de frecuencias relativas (regla 1) de un suceso, tiende a aproximarse a la probabilidad real.

La ley de los grandes números indica que los estimados por frecuencias relativas de la regla 1 tienden a mejorar si se hacen más observaciones. Esta ley refleja una simple noción fundamentada en el sentido común: un estimado de probabilidad basado en sólo unos cuantos ensayos puede desviarse en cantidades sustanciales; pero, con un número muy grande de ensayos, el estimado tiende a ser mucho más preciso. Por ejemplo, es muy fácil que una encuesta de opinión entre sólo una docena de personas seleccionadas al azar resulte errónea en gran medida, pero si se aplica a miles de personas seleccionadas al azar, puede acercarse bastante a los valores reales de la población.

EJEMPLO Volando alto Calcule la probabilidad de que un adulto que se seleccionó aletoriamente haya volado en una línea aérea comercial.

SOLUCIÓN El espacio muestral consiste en dos sucesos simples: la persona que se seleccionó ya voló en una línea aérea comercial o no lo ha hecho. Puesto que el espacio muestral consiste en sucesos que no son igualmente probables, no es posible utilizar el método clásico (regla 2). Con estos resultados de una encuesta de Gallup, podemos usar el método de frecuencias relativas (regla 1): de 855 adultos que se seleccionaron al azar, 710 indicaron que ya volaron en líneas aéreas comerciales. Obtenemos el resultado siguiente:

$$P(\text{haber volado en una línea aérea comercial}) = \frac{710}{855} \approx 0.830$$

EJEMPLO Ruleta Usted planea apostar al número 13 en el próximo giro de una ruleta. ¿Cuál es la probabilidad de que pierda?

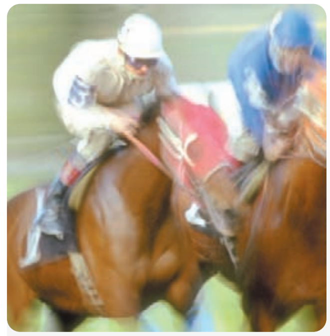
SOLUCIÓN Una ruleta tiene 38 ranuras distintas y sólo una corresponde al número 13. La ruleta se diseñó de manera que las 38 ranuras sean igualmente probables de resultar. De las 38 ranuras, 37 resultan en una pérdida. Ya que el espacio muestral incluye resultados igualmente probables, usamos el método clásico (regla 2) para obtener

$$P(\text{pérdida}) = \frac{37}{38}$$

EJEMPLO Choque de meteoritos ¿Cuál es la probabilidad de que su automóvil sea impactado por un meteorito este año?

SOLUCIÓN La ausencia de datos históricos de meteoritos que chocan contra automóviles impide usar el método de frecuencias relativas de la regla 1. Hay dos posibles resultados (chocar o no chocar), pero no son igualmente probables, de tal forma que no podemos usar el método clásico de la regla 2. Esto nos deja con la regla 3, por medio de la cual hacemos un estimado subjetivo. En tal caso, todos sabemos que la probabilidad en cuestión es muy, muy pequeña. Estimemos que sea, digamos, de 0.000000000001 (equivalente a una en un billón). Este estimado subjetivo, que se basa en nuestro conocimiento general, puede encontrarse en el campo general de la probabilidad real.

En problemas de probabilidad básicos del tipo de los que estamos considerando, es muy importante examinar con cuidado la información disponible e identificar correctamente el número total de posibles resultados. En algunos casos, el número total de resultados posibles está dado, pero en otros tiene que calcularse, como en el siguiente ejemplo, que requiere que calculemos el número total de resultados posibles.



Probabilidades subjetivas en el hipódromo

Algunos investigadores han estudiado la capacidad para estimar probabilidades subjetivas realistas de los apostadores en los hipódromos. (Véase "Racetrack Betting: Do Bettors Understand the Odds?", de Brown, D'Amato y Gertner, revista *Chance*, vol. 7, núm. 3). Después de analizar los resultados de 4400 carreras, los autores concluyeron que, aunque los apostadores sobreestiman un poco las probabilidades de ganar de los que no son favoritos y subestiman ligeramente las probabilidades de ganar de los favoritos, su desempeño general es muy bueno. Las probabilidades subjetivas se calcularon a partir de las ganancias de los apostadores, con base en las cantidades que se apostaron, en tanto que las probabilidades reales se calcularon a partir de los resultados reales de las carreras.

EJEMPLO Pena de muerte Se seleccionan adultos al azar para una encuesta de Gallup; a ellos se les pregunta si están a favor de la pena de muerte para una persona convicta por homicidio. Las respuestas incluyen a 319 personas que están a favor de la pena de muerte, 133 personas que están en contra y 39 que no tienen una opinión al respecto. Con base en tales resultados, estime la probabilidad de que una persona seleccionada aleatoriamente esté a favor de la pena de muerte.

SOLUCIÓN *Sugerencia:* En lugar de tratar de formular una respuesta directamente del extracto escrito, resuma la información dada en un formato que le permita comprenderla mejor. Por ejemplo:

319	a favor de la pena de muerte
133	en contra de la pena de muerte
39	sin opinión
491	total

Ahora utilicemos el método de frecuencias relativas (regla 1) como sigue:

$$P\text{ (personas a favor de la pena de muerte)}$$

$$= \frac{\text{número de personas a favor de la pena de muerte}}{\text{total}} = \frac{319}{491} = 0.650$$

Estimamos que hay una probabilidad de 0.650 de que cuando un adulto se selecciona al azar, él o ella estén a favor de la pena de muerte para alguien convicto por homicidio. Como sucede con todas las encuestas, la precisión de tal resultado depende de la calidad del método de muestreo y del procedimiento de la encuesta. Ya que la encuesta fue realizada por la organización Gallup, es probable que los resultados sean razonablemente precisos. El capítulo 6 incluirá procedimientos más avanzados para analizar resultados de encuesta como éstos.

1ro 2do 3ro	
niño–niño–niño	
niño–niño–niña	→ exactamente 2 niños
niño–niña–niño	→ exactamente 2 niños
niño–niña–niña	
niña–niño–niño	→ exactamente 2 niños
niña–niño–niña	
niña–niña–niño	
niña–niña–niña	

EJEMPLO Género de hijos Determine la probabilidad de que una pareja con tres hijos tenga exactamente dos niños. Suponga que es igualmente probable dar a luz un niño que una niña y que el género de cualquier hijo no influye en el género del otro.

SOLUCIÓN El mayor obstáculo es identificar correctamente el espacio muestral. Esto implica más que trabajar sólo con los números 2 y 3, que se dieron en el planteamiento del problema. El espacio muestral consiste en ocho diferentes formas en que los tres hijos pueden presentarse; las listamos al margen. Como los ocho resultados son igualmente probables, utilizamos la regla 2. De los ocho posibles resultados, tres corresponden exactamente a dos niños, así que

$$P(2 \text{ niños en } 3 \text{ nacimientos}) = \frac{3}{8} = 0.375$$

INTERPRETACIÓN Existe una probabilidad de 0.375 de que si un matrimonio tiene tres hijos, exactamente dos de ellos sean niños.

Los enunciados de las tres reglas para calcular probabilidades y los ejemplos anteriores parecen sugerir que siempre debemos usar la regla 2 cuando un procedimiento tiene resultados igualmente probables. En realidad, muchos procedimientos son tan complicados que el uso del método clásico (regla 2) resulta impráctico. En el juego del Solitario, por ejemplo, los resultados (del reparto) son todos igualmente probables, pero es muy frustrante tratar de utilizar la regla 2 para calcular la probabilidad de ganar. En estos casos es posible obtener buenos estimados con mayor facilidad usando el método de frecuencias relativas (regla 1). Es muy común que las simulaciones sean útiles cuando se emplea este método. (Una **simulación** de un procedimiento es un proceso que se comporta en las mismas formas que el procedimiento mismo; por lo tanto, produce resultados similares). Por ejemplo, al estimar la probabilidad de ganar en el Solitario, es mucho más fácil usar la regla 1 y repetir el juego muchas veces (o correr una simulación por computadora) que realizar los cálculos extremadamente complejos que se requieren con la regla 2.

EJEMPLO Día de Acción de Gracias Si se selecciona un año al azar, calcule la probabilidad de que el Día de Acción de Gracias sea un *a*) miércoles, *b*) jueves.

SOLUCIÓN

a. El Día de Acción de Gracias se celebra siempre el cuarto jueves de noviembre. Por lo tanto, es imposible que un Día de Acción de Gracias caiga en miércoles. Cuando un suceso es imposible, decimos que su probabilidad es 0.

b. Es cierto que el Día de Acción de Gracias es un jueves. Cuando es seguro que un suceso ocurra, decimos que su probabilidad es 1.

Ya que cualquier suceso imaginable es imposible, o cierto, o está en alguna parte intermedia, se deduce que la probabilidad matemática de cualquier suceso es 0, 1, o un número entre 0 y 1 (véase figura 3-2).

- **La probabilidad de un suceso imposible es 0.**
- **La probabilidad de un suceso que ocurrirá con certeza es 1.**
- $0 \leq P(A) \leq 1$ **para cualquier suceso *A*.**

En la figura 3-2, la escala de 0 hasta 1 se muestra a la izquierda, mientras que las expresiones más comunes y familiares de probabilidad se indican a la derecha.

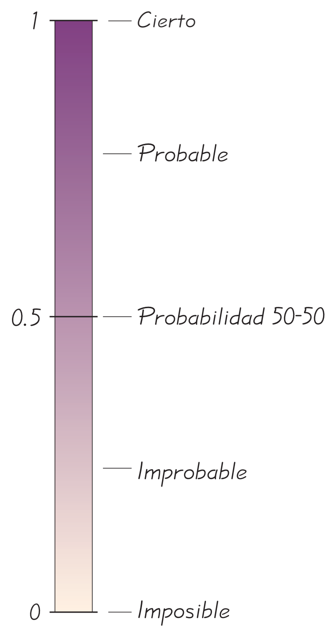


FIGURA 3-2 Valores posibles para probabilidades

Sucesos complementarios

Algunas veces necesitamos calcular la probabilidad de que un suceso *A* *no* ocurra.

Definición

Complemento de un suceso *A*, denotado por $\overline{A}$: consiste en todos los resultados en los cuales el suceso *A* *no* ocurre.

EJEMPLO Género al nacer En realidad nacen más niños que niñas. En un grupo típico, hay 205 bebés recién nacidos y 105 de ellos son niños. Si un bebé del grupo es seleccionado al azar, ¿cuál es la probabilidad de que el bebé *no* sea un niño?

continúa

¿Qué tan probable?

¿Cómo interpretamos términos como probable, improbable o extremadamente probable? La Federal Aviation Administration (FAA) interpreta estos términos de la manera siguiente: *Probable:* una probabilidad del orden de 0.00001 o mayor para cada hora de vuelo. Se espera que ocurran sucesos de este tipo varias veces durante la vida operacional de cada aeroplano. *Improbable:* una probabilidad del orden de 0.00001 o menor. Sucesos de esta clase no se espera que ocurran durante toda la vida operacional de un solo aeroplano de un tipo en particular, aunque pueden ocurrir durante toda la vida operacional de todos los aeroplanos de un tipo en particular. *Extremadamente improbable:* una probabilidad en el orden de 0.000000001 o menor. Eventos como éste son tan improbables que no es necesario considerar su ocurrencia.

SOLUCIÓN Ya que 105 de los 205 bebés son niños, se deduce que 100 de ellos son niñas, entonces

$$P(\text{no seleccionar a un niño}) = P(\overline{\text{niño}}) = P(\text{niña}) = \frac{100}{205} = 0.488$$

Aun cuando es difícil desarrollar una regla universal para el redondeo de probabilidades, el siguiente lineamiento se aplicará a la mayoría de los problemas en este texto.

Redondeo de probabilidades

Cuando se expresa el valor de una probabilidad, hay que dar la fracción o el número decimal *exactos*, o redondear los resultados decimales finales a tres cifras significativas. (*Sugerencia:* Cuando una probabilidad no sea una fracción simple como $2/3$ o $5/9$, exprésela como decimal para que el número resulte más claro).

Todos los dígitos en un número son significativos, excepto los ceros, que se incluyen para la colocación apropiada del punto decimal.

EJEMPLOS

- La probabilidad de 0.021491 tiene cinco dígitos relevantes (21491), por lo cual puede redondearse a 0.0215, con tres dígitos relevantes.
- La probabilidad de $1/3$ puede permanecer como fracción o redondearse a 0.333. *No* redondee a 0.3.
- La probabilidad de caras en un lanzamiento de una moneda puede expresarse como $1/2$ o 0.5; ya que 0.5 es exacto, no hay necesidad de expresarlo como 0.500.
- La fracción $432/7842$ es exacta, pero su valor no es evidente. Exprésela como el decimal 0.0551.

La expresión matemática de la probabilidad como un número entre 0 y 1 es un concepto importante en esta sección. Esta forma de expresión es fundamental y común en los procedimientos estadísticos; la usaremos de aquí en adelante en. Por ejemplo, un resultado de computadora típico puede incluir una expresión "valor *P*" como "nivel de significancia 0.001". Más tarde analizaremos el significado del valor *P*, pero tales valores son esencialmente probabilidades del tipo que se analizó en esta sección. Por ahora, usted debe reconocer que una probabilidad de 0.001 (equivalente a $1/1000$) corresponde a un suceso tan infrecuente, que puede ocurrir un promedio de sólo una vez en mil ensayos.

Posibilidades

Las expresiones de probabilidad a veces se proponen como *posibilidades*, como 50:1 (o "50 a 1"). Una grave desventaja de las posibilidades es que hacen que muchos cálculos sean extremadamente difíciles. Por ello, los estadísticos, los matemáticos y los científicos prefieren usar probabilidades. La ventaja de las posibilidades es que facilitan el manejo de las transacciones de dinero asociadas con los juegos de azar, por lo cual tienden a usarse en casinos, loterías e hipódromos. Note que, en las tres definiciones siguientes, las posibilidades reales en contra y las posibilidades reales a favor describen la probabilidad real de algún suceso, pero las posibilidades de pago describen la relación entre la apuesta y la cantidad del pago. Las pistas de carreras y los casinos están en el negocio con la finalidad de lograr su propio beneficio. Por ello, las posibilidades de pago no serán las mismas que las posibilidades reales.

Definición

Las **posibilidades reales en contra** de que ocurra un suceso A son el cociente $P(\overline{A})/P(A)$, casi siempre expresado en la forma $a{:}b$ (o "a a b"), donde a y b son enteros que no tienen factores comunes.

Las **posibilidades reales a favor** del suceso A son el recíproco de las posibilidades reales en contra de ese suceso. Si las posibilidades en contra de A son $a{:}b$, entonces las posibilidades a favor de A son $b{:}a$.

Las **posibilidades de pago** contra el suceso A representan la proporción de la ganancia neta (si usted gana) con respecto a la cantidad de la apuesta.

posibilidades de pago en contra del suceso A = (ganancia neta):(cantidad apostada)

EJEMPLO Si usted apuesta \$5 al número 13 en la ruleta, su probabilidad de ganar es $1/38$, en tanto que las posibilidades de pago están dadas por el casino como 35:1.

a. Calcule las posibilidades reales en contra del resultado de 13.

b. ¿Cuánta ganancia neta podría usted obtener si gana apostando al 13?

c. Si el casino estuviera funcionando solamente por diversión y las posibilidades de pago fueran cambiadas para igualar las posibilidades reales en contra del 13, ¿cuanto ganaría usted si el resultado fuera 13?

SOLUCIÓN

a. Con $P(13) = 1/38$ y $P(\text{no } 13) = 37/38$, tenemos

$$\text{posibilidades reales en contra del 13} = \frac{P(\text{no } 13)}{P(13)} = \frac{37/38}{1/38} = \frac{37}{1} \text{ o } 37{:}1$$

b. Puesto que las posibilidades de pago en contra del 13 son 35:1, tenemos

$$35{:}1 = (\text{ganancia neta}){:}(\text{monto apostado});$$

entonces, hay una ganancia de \$35 por cada \$1 que se apuesta. Para una apuesta de \$5, la ganancia neta es de \$175. El apostador que gane podría recoger \$175 más la apuesta original de \$5.

c. Si el casino estuviera funcionando por diversión y no por ganancia, las posibilidades de pago serían iguales a las posibilidades reales en contra del resultado de 13. Si las posibilidades de pago se cambiaran de 35:1 a 37:1, usted obtendría una ganancia neta de \$37 por cada \$1 que apostara. Si usted apuesta \$5, su ganancia neta sería de \$185. (El casino logra su ganancia pagando sólo \$175, en lugar de los \$185 que se pagarían con un juego de ruleta justo, en lugar de uno que favorece al casino).

Puedes apostarlo

En una lotería estatal típica, la "casa" tiene una ventaja del 65 o el 70%, ya que sólo entre el 35 y el 40% del dinero que se apuesta se devuelve en forma de premios. La ventaja de la casa en los hipódromos suele ser de alrededor del 15%. En los casinos, la ventaja de la casa es del 5.26% para la ruleta, del 5.9% para el 21, del 1.4% para los dados y del 3 al 22% para las máquinas tragamonedas. Algunos jugadores profesionales pueden ganar sistemáticamente en el 21, usando complicadas técnicas de conteo de cartas. Ellos saben cuándo un mazo tiene una proporción elevada de cartas altas; es entonces cuando hacen apuestas cuantiosas. Muchos casinos reaccionan expulsando a los contadores de cartas o revolviendo los mazos con más frecuencia.

Probabilidades que desafían a la intuición

En ciertos casos, nuestros estimados subjetivos de valores de probabilidad son drásticamente distintos de las probabilidades reales. He aquí un ejemplo clásico: si usted inhala profundamente, hay una probabilidad mayor al 99% de que inhale una molécula que César exhaló en el último aliento al morir. En este mismo ánimo morboso y poco intuitivo, si la fatal taza con cicuta que mató a Sócrates hubiera contenido en su mayor parte agua, el siguiente vaso de agua que usted beba muy probablemente contendrá una de esas mismas moléculas. He aquí un ejemplo más, pero menos morboso, que puede verificar: en grupos de 25 estudiantes, la probabilidad de que al menos dos cumplan años el mismo día es de más del 50%.

3-2 Destrezas y conceptos básicos

En los ejercicios 1 y 2 exprese el grado indicado de probabilidad como un valor de probabilidad.

1. Identificación de valores de probabilidad
a. "Usted tiene una probabilidad de 50-50 de escoger el camino correcto".
b. "Hay un 20% de probabilidad de que llueva mañana".
c. "Usted tiene una probabilidad de un pelo de rana de casarse con mi hija".

2. Identificación de valores de probabilidad
a. "Hay un 90% de probabilidad de que mañana nieve".
b. "Definitivamente, por la noche oscurecerá".
c. "Usted tiene una probabilidad en diez de estar en lo correcto".

3. Identificación de valores de probabilidad ¿Cuáles de los siguientes valores *no pueden* ser probabilidades?

$$0, \quad 1, \quad -1, \quad 2, \quad 0.0123, \quad 3/5, \quad 5/3, \quad \sqrt{2}$$

4. Identificación de valores de probabilidad
a. ¿Cuál es la probabilidad de que ocurra un suceso inevitable?
b. ¿Cuál es la probabilidad de un suceso imposible?
c. Un espacio muestral consiste en 10 sucesos separados que son igualmente probables. ¿Cuál es la probabilidad de cada uno?
d. En un examen de verdadero/falso, ¿cuál es la probabilidad de responder una pregunta correctamente si usted elige al azar?
e. En un examen de opción múltiple, con cinco posibles respuestas para cada pregunta, ¿cuál es la probabilidad de responder una pregunta correctamente si usted elige al azar?

5. Género de hijos En esta sección, dimos un ejemplo que incluye una lista de los ocho resultados posibles cuando una pareja tiene tres hijos. Remítase a esa lista y calcule la probabilidad de cada suceso.
a. De entre tres hijos, hay exactamente una niña.
b. De entre tres hijos, hay exactamente dos niñas.
c. De entre tres hijos, todos son niñas.

6. Teléfonos celulares y cáncer cerebral En un estudio de 420,000 usuarios de teléfono celular en Dinamarca, se encontró que 135 desarrollaron cáncer cerebral o del sistema nervioso. Estime la probabilidad de que un usuario de teléfono celular que se seleccionó al azar desarrolle un cáncer de este tipo. Se encontró que la probabilidad para la población general es de 0.000340; ¿es el resultado muy diferente de éste? ¿Qué sugiere el resultado acerca de los teléfonos celulares como causantes de cáncer de este tipo, como ya se afirmó?

7. Probabilidad de un jonrón El jugador de beisbol Barry Bonds rompió un récord importante cuando dio 73 jonrones en la temporada 2001. Durante esa temporada, estuvo al bat 476 veces. Si se selecciona al azar una de las ocasiones que estuvo al bat, calcule la probabilidad de que sea una de las veces que pegó un jonrón. ¿Difiere mucho el resultado de la probabilidad de 0.0715 que resulta de sus 567 jonrones en 7932 ocasiones que estuvo al bat?

8. Ser alcanzado por un rayo En un año reciente, de los 281,421,906 habitantes de Estados Unidos, 389 fueron alcanzados por un rayo. Calcule la probabilidad de que a una persona que se selecciona al azar en Estados Unidos sea alcanzada por un rayo este año.

Uso de la probabilidad para identificar sucesos infrecuentes. *En los ejercicios 9 a 16 considere un suceso como "infrecuente" si su probabilidad es igual o menor que 0.05. (Esto equivale al mismo criterio que se usa comúnmente en estadística inferencial, pero* *como el valor de 0.05 no es absolutamente rígido, algunas veces se emplean otros valores en su lugar; por ejemplo, 0.01).*

9. Probabilidad de un resultado equivocado La tabla 3-1 muestra que, de 85 mujeres embarazadas, la prueba de embarazo arrojó un resultado equivocado cinco veces.
- **a.** Con base en los resultados disponibles, calcule la probabilidad de obtener un resultado equivocado de la prueba para una mujer embarazada.
- **b.** Para mujeres no embarazadas, ¿será "infrecuente" que el resultado de la prueba esté equivocado?

10. Probabilidad de un resultado equivocado La tabla 3-1 muestra que de 14 mujeres que *no* están embarazadas, la prueba de embarazo produjo un resultado equivocado tres veces.
- **a.** Con base en los resultados disponibles, calcule la probabilidad de obtener un resultado de prueba equivocado para una mujer no embarazada.
- **b.** Para mujeres no embarazadas ¿será "infrecuente" que el resultado de la prueba esté equivocado?

11. Encuesta de tabaquismo En una encuesta de Gallup, se interrogó a 1038 adultos acerca de los efectos del tabaquismo pasivo; 52 de ellos indicaron que tales efectos "no son dañinos en absoluto".
- **a.** Si usted selecciona al azar a uno de los adultos que se encuestaron, ¿cuál es la probabilidad de seleccionar a alguien que opine que ser fumador pasivo no es dañino en absoluto?
- **b.** ¿Es "infrecuente" que alguien opine que ser fumador pasivo no es dañino en absoluto?

12. Fármaco reductor del colesterol En un ensayo clínico de Lipitor, un fármaco común que se usa para disminuir el colesterol, a un grupo de pacientes se les administró un tratamiento de tabletas de Atorvastatin de 10 miligramos. En dicho grupo, 19 pacientes sufrieron síntomas de gripe y 844 no los sufrieron (según datos de Pfizer, Inc.).
- **a.** Estimar la probabilidad de que un paciente que toma el fármaco sufra síntomas de gripe.
- **b.** ¿Es "infrecuente" que un paciente que toma el fármaco sufra síntomas de gripe?

13. Pasajeros de líneas aéreas "rebotados" En un año reciente, a 2624 pasajeros de American Airlines se les impidió abordar sus vuelos contra su voluntad, en tanto que hubo otros 168,262 que fueron "rebotados" voluntariamente a cambio de efectivo o vales.
- **a.** Estime la probabilidad de que un pasajero rebotado de American Airlines, que se selecciona al azar, sea uno de los que fueron rebotados contra su voluntad.
- **b.** ¿Es "infrecuente" que alguien sea "rebotado" en contra de su voluntad?

14. Llegadas de vuelo a tiempo Un estudio de 150 vuelos de American Airlines, seleccionados aleatoriamente, mostró que 108 llegaron a tiempo (según datos del Departamento del Transporte de Estados Unidos).
- **a.** ¿Cuál es la probabilidad estimada de que un vuelo de American Airlines llegue *retrasado*?
- **b.** ¿Es "infrecuente" que un vuelo de American Airlines llegue retrasado?

15. Adivinación de fechas del nacimiento En su primera cita, Kelly le pide a Mike que adivine su fecha de nacimiento, omitiendo el año.
- **a.** ¿Cuál es la probabilidad de que Mike adivine correctamente? (Ignore los años bisiestos).
- **b.** ¿Sería "infrecuente" que él adivinara con acierto en el primer intento?
- **c.** Si usted fuera Kelly, y Mike adivinara correctamente en su primer intento, ¿creería que él tuvo un golpe de suerte o estaría convencida de que él ya sabía la fecha en que usted nació?
- **d.** Si Kelly le pide a Mike que adivine su edad, y la respuesta de Mike es más alta por 15 años, ¿cuál es la probabilidad de que Mike y Kelly tengan una segunda cita?

16. Lotería En la antigua lotería del estado de Nueva York, usted tenía que escoger seis números entre 1 y 54, inclusive. Había 25,827,165 diferentes combinaciones de seis números posibles y se tenía que seleccionar la combinación correcta de los seis números para ganar el premio mayor. Para una apuesta de $1, usted escogía dos distintas combinaciones de seis números. (No era posible seleccionar sólo una combinación de seis números, usted tenía que seleccionar dos).

a. Si usted apostaba $1 y seleccionaba dos combinaciones diferentes de seis números, ¿cuál era la probabilidad de ganar el premio mayor?

b. ¿Era infrecuente ganar el premio mayor?

17. Probabilidad de un cumpleaños

a. Si se selecciona una persona al azar, calcule la probabilidad de que su cumpleaños sea el 18 de octubre, que es el Día Nacional de la Estadística en Japón. Ignore los años bisiestos.

b. Si se selecciona a una persona aleatoriamente, calcule la probabilidad de que su cumpleaños sea cualquier otro día. Ignore los años bisiestos.

c. Seleccione al azar a una persona y calcule la probabilidad de que naciera un día de la semana que termine con la letra *s* o con la letra *o*.

18. Probabilidad de reconocimiento de marca

a. En un estudio de reconocimiento de marcas, 831 consumidores conocían la sopa Campbell's y 18 no (según datos de Total Research Corporation). Use dichos resultados para estimar la probabilidad de que un consumidor que se selecciona al azar reconozca la sopa Campbell's.

b. *Estime* la probabilidad de que un consumidor adulto estadounidense, seleccionado al azar, conozca el nombre de la marca McDonald's, la cadena de restaurantes de comida rápida más famosa en Estados Unidos.

c. *Estime* la probabilidad de que un consumidor adulto estadounidense que se selecciona aleatoriamente reconozca el nombre de la marca Veeco Instruments, un fabricante de productos de microelectrónica.

19. Encuesta del pastel de frutas (*fruitcake*) En una encuesta de Bruskin-Goldring Research, se preguntó cómo debía usarse un pastel de frutas. Ciento treinta y dos personas respondieron que como tope para una puerta y 880 citaron otros fines, incluyendo como comida para pájaros, relleno para terrenos y regalo. Si se selecciona al azar a una de estas personas, ¿cuál es la probabilidad de que sea alguien que usaría el pastel de frutas como tope para una puerta?

20. Probabilidad de un accidente automovilístico De entre 400 conductores aleatoriamente seleccionados, en el rango de edades de 20 a 24 años, 136 sufrieron un accidente automovilístico durante el año anterior (de acuerdo con datos del Consejo de Seguridad Nacional de Estados Unuidos). Si se selecciona al azar a un conductor en ese rango de edad, ¿cuál es la probabilidad aproximada de que él, o ella, sufra un accidente automovilístico durante el año próximo? ¿Será el valor resultante suficientemente alto como para preocupar a los individuos de 20 a 24 años de edad?

21. Probabilidad de ganar en el Solitario Remítase al conjunto de datos 27 del Apéndice B y suponga que se juega el mismo Solitario de Microsoft.

a. Estime la probabilidad de ganar cuando se juega una partida.

b. Estime la probabilidad de ganar $208 agotando todas las cartas.

22. Probabilidad de reacción adversa a un fármaco Cuando el fármaco Viagra se probó clínicamente, 117 pacientes reportaron dolor de cabeza y 617 no lo hicieron (según datos de Pfizer, Inc.). Use esta muestra para estimar la probabilidad de que un usuario de Viagra sufra dolor de cabeza. ¿La probabilidad es suficientemente alta como para preocupar a los usuarios de Viagra?

23. Género de hijos: construcción de espacio muestral La sección 3-2 incluye una tabla que resume los resultados de género para una pareja que planea tener tres hijos.

a. Construya una tabla similar para una pareja que planea tener dos hijos.

b. Suponiendo que los resultados que se listan en el inciso *a*) sean igualmente probables, calcule la probabilidad de tener dos niñas.
c. Calcule la probabilidad de tener exactamente un hijo de cada género.

24. Genética: construcción de espacio muestral Ambos padres tienen el par de genes de color de ojos café/azul, y cada uno transmite un gene a un hijo. Suponga que si el hijo tiene al menos un gene café, ese color dominará y los ojos serán cafés. (La determinación real del color de los ojos es un tanto más complicada).
a. Haga una lista de los posibles resultados diferentes. Suponga que dichos resultados son igualmente probables.
b. ¿Cuál es la probabilidad de que un hijo de estos padres tenga el par de genes azul/azul?
c. ¿Cuál es la probabilidad de que el hijo tenga ojos cafés?

25. Posibilidades en el Derby de Kentucky Cuando el caballo Monarchos ganó el 127º Derby de Kentucky, una apuesta de $2 a que Monarchos ganaría resultó en un reintegro de $23.
a. ¿Qué ganancia neta hubo al ganar con una apuesta de $2 a Monarchos?
b. ¿Cuáles fueron las posibilidades de pago en contra de que Monarchos ganara?
c. Con base en el paseo preliminar a la carrera, los apostadores colectivamente creyeron que Monarchos tenía una probabilidad de ganar de 1/15. Suponiendo que 1/15 era la probabilidad real de la victoria de Monarchos, ¿cuáles fueron las posibilidades reales en contra?
d. Si las posibilidades de pago fueran iguales a las posibilidades reales que se calcularon en el inciso *c*), ¿cuanto valdría un boleto de $2 después de que Monarchos ganó?

26. Cálculo de posibilidades en la ruleta Una rueda de ruleta tiene 38 ranuras, una ranura es 0, otra es 00 y las otras están numeradas del 1 hasta el 36. Usted apuesta a número impar.
a. ¿Cuál es su probabilidad de ganar?
b. ¿Cuáles son las posibilidades reales en contra?
c. Cuando se apuesta a número impar, las posibilidades de pago son 1:1. ¿Qué ganancia obtendría usted si apostara $18, si pudiera, de alguna manera, convencer al casino de cambiar sus posibilidades de pago para que fueran las mismas que las posibilidades reales en contra? (*Sugerencia:* No trate realmente de convencer a ningún casino de esto; su sentido del humor está ausente por completo cuando se trata de asuntos como éste).

3-2 Más allá de lo básico

27. Interpretación de la eficacia Se diseña un experimento doble ciego para probar la eficacia del fármaco Estadisticzeno, como tratamiento para la ceguera ante los números. Cuando a los sujetos se les trata con Estadisticzeno, parecen mostrar mejoría. Los investigadores calculan que hay una probabilidad de 0.04 de que el grupo de tratamiento muestre mejoría si el fármaco no surte efecto. ¿Qué concluye acerca de la eficacia del Estadisticzeno?

28. Determinación de un jurado aleatorio Un abogado defiende a un cliente a quien se le acusó de no cumplir con sus obligaciones de pensión alimenticia. La junta de jurados potenciales consta de 20 mujeres, por lo que el abogado calcula que hay una probabilidad de 1/1,048,576 de que 20 personas seleccionadas aleatoriamente sean todas mujeres. ¿Hay fundamento para argumentar que la junta de jurados es injusta para su cliente?

29. Cálculo de probabilidad a partir de posibilidades Puesto que las posibilidades reales en contra de un suceso A sean $a{:}b$, entonces $P(A) = b/(a + b)$. Calcule la probabilidad de que Millenium gane su próxima carrera, ya que las posibilidades reales en contra son de 3:5.

30. Riesgo relativo y cociente de posibilidades En un ensayo clínico de 734 sujetos, a quienes se les trató con Viagra, 117 reportaron dolor de cabeza. En un grupo control con 725 sujetos sin dicho tratamiento, 29 reportaron dolores de cabeza. Si la proporción de dolores de cabeza en el grupo de tratamiento se denota como p_t y la proporción de dolores de cabeza en el grupo control como p_c, el **riesgo relativo** es p_t/p_c. El riesgo relativo es una medida de la fuerza del efecto del tratamiento con Viagra. Otra medida como ésta es el **cociente de posibilidades**, que es el cociente de las posibilidades a favor de un dolor de cabeza en el grupo de tratamiento, entre las posibilidades a favor de un dolor de cabeza en el grupo control, el cual se calcula evaluando lo siguiente:

$$\frac{p_t/(1-p_t)}{p_c/(1-p_c)}$$

El riesgo relativo y el cociente de posibilidades a menudo se utilizan en estudios médicos y epidemiológicos. Calcule el riesgo relativo y el cociente de posibilidades para los datos del dolor de cabeza.

31. Años bisiestos y adivinación de cumpleaños En el inciso *a*) del ejercicio 17, para encontrar la probabilidad de que una persona que se selecciona al azar cumpla años el 18 de octubre, se ignoraron los años bisiestos.

a. Recalcule tal probabilidad considerando que un año bisiesto ocurre cada cuatro años. (Exprese su respuesta en forma de fracción exacta).

b. Los años bisiestos ocurren en años divisibles entre cuatro, salvo en los años centesimales (años que terminan en 00), en los que ocurre un año bisiesto en una serie de cuatro. Por ejemplo, los años 1700, 1800, y 1900 no fueron años bisiestos, mientras que el 2000 sí lo fue. Calcule la probabilidad exacta para el caso que se menciona, incluyendo los años bisiestos, y exprésela como una fracción exacta.

32. Moscas en una naranja Si dos moscas se paran en una naranja, calcule la probabilidad de que ambas se localicen en puntos pertenecientes al mismo hemisferio.

33. Puntos en un palo En un palo recto se seleccionan al azar dos puntos longitudinales. Después, se rompe el palo en esos dos puntos. Calcule la probabilidad de que los tres pedazos que quedan se puedan acomodar para formar un triángulo. (Quizás éste sea el ejercicio más difícil del libro).

3-3 Regla de la suma

El objetivo principal de esta sección es introducir la *regla de la suma* como un método para calcular probabilidades que pueden expresarse de la forma $P(A \text{ o } B)$, o sea, la probabilidad de que ocurra el suceso A o de que ocurra el suceso B (o de que ambos ocurran), como único resultado de un procedimiento. La palabra clave que debemos recordar es *o*. En todo este texto usaremos la expresión *o inclusive*, que significa: uno o el otro, o ambos. (Con la excepción del ejercicio 27, no consideramos el *o exclusivo*, que significa que o bien uno o bien el otro, pero no ambos).

En la sección anterior presentamos aspectos fundamentales de la probabilidad y estudiamos sucesos calificados como *simples*. En esta sección y en la siguiente, introducimos *sucesos compuestos*.

Definición

Suceso compuesto: cualquier suceso que combina dos o más sucesos simples.

Notación de la regla de la suma

$P(A \text{ o } B) = P$(ocurre el suceso A u ocurre el suceso B o ambos ocurren).

Las probabilidades juegan un papel prominente en la genética. Observe la figura 3-3, que presenta una muestra de chícharos, como los que usó Mendel en sus famosos experimentos de hibridación. Los chícharos que se muestran tienen vainas verdes o amarillas y flores moradas o blancas. En esta muestra de 14 chícharos, ¿cuántos de ellos tienen "vainas verdes o flores moradas"? (Recuerde, "vaina verde o flor morada" significa realmente "vaina verde, o flor morada, o ambas"). La revisión de la figura 3-3 debe indicar que, en total, 12 chícharos tienen vainas verdes o flores moradas. (*Nota importante: Es erróneo* sumar los ocho chícharos con vainas verdes a los nueve chícharos con flores moradas, ya que el total de 17 toma en cuenta dos veces a cinco de los chícharos, pero éstos son individuos, por lo cual deben contarse una vez cada uno). Puesto que 12 de los 14 chícharos tienen "o vainas verdes o flores moradas", la probabilidad de seleccionar un chícharo al azar con una vaina verde o una flor morada se expresa como P(vaina verde o flor morada) $= 12/14 = 6/7$.

El ejemplo sugiere una regla general por medio de la cual sumamos el número de resultados que corresponden a cada uno de los sucesos en cuestión:

Para calcular la probabilidad de que un suceso *A* ocurra o un suceso *B* ocurra, calcule el número total de formas en que *A* puede ocurrir y el número de formas en que *B* puede ocurrir, pero *calcule ese total de tal manera que ningún resultado se cuente más de una vez.*

Un método consiste en combinar el número de formas en que un suceso A puede ocurrir con el número de formas en que un suceso B puede ocurrir y, si hay cualquier traslape entre estos dos conjuntos, compensar restando el número de resultados que se contaron dos veces, como se hace en la regla siguiente:

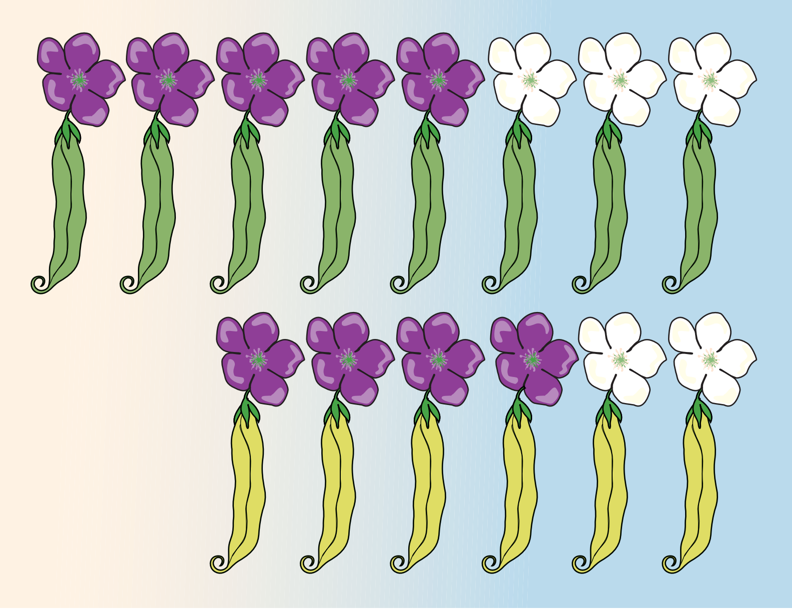


FIGURA 3-3 Chícharos utilizados en un estudio genético

Regla formal de la suma

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

donde $P(A \text{ y } B)$ denota la probabilidad de que A y B ocurran al mismo tiempo, como resultado de ensayo o de procedimiento.

Aunque la regla formal de la suma se presenta como una fórmula, en general es mejor entender el espíritu de la regla y aplicarlo intuitivamente de la siguiente forma:

Regla intuitiva de la suma

Para obtener P(A o B), calcule la suma del número de formas en que puede ocurrir el suceso A y el número de formas en que puede ocurrir el suceso B, *sumando de tal modo que cada resultado se cuente sólo una vez*. P(A o B) es igual a esa suma dividida entre el número total de resultados en el espacio muestral.

En la figura 3-4 se muestra un diagrama de Venn, que nos ofrece una comprensión visual de la regla formal de la suma. En esta figura se observa que la probabilidad de A o B es igual a la probabilidad de A (círculo izquierdo), más la probabilidad de B (círculo derecho) menos la probabilidad de *A y B* (región media con forma de balón de futbol americano). La figura nos muestra que la suma de las áreas de los dos círculos haría que se contara dos veces la región media. Éste es el concepto básico que sustenta la regla de la suma. Por la relación entre la regla de la suma y el diagrama de Venn que se muestra en la figura 3-4, es común el uso de la notación $P(A \cup \text{B})$ en lugar de $P(A \text{ o } B)$. Asimismo, se utiliza con frecuencia la notación $P(A \cap B)$ en lugar de $P(A \text{ y } B)$, de modo que la regla formal de la suma se expresa como

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

La regla de la suma se simplifica si A y B no pueden ocurrir simultáneamente, porque $P(A \text{ y } B)$ se convierte en cero. La figura 3-5 ilustra que si A y B no se traslapan, tenemos $P(A \text{ o } B) = P(A) + P(B)$. La definición siguiente formaliza la ausencia de intercepto que se muestra en la figura 3-5.

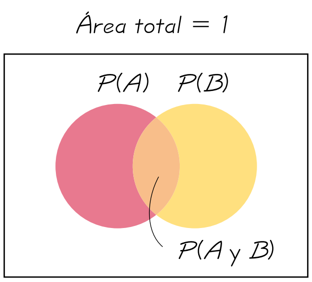


FIGURA 3-4 Diagrama de Venn que muestra sucesos traslapados

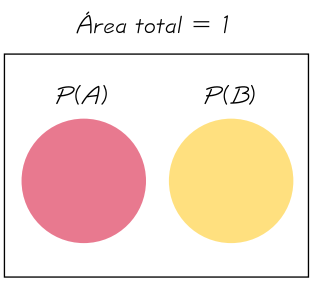


FIGURA 3-5 Diagrama de Venn que muestra sucesos no traslapados

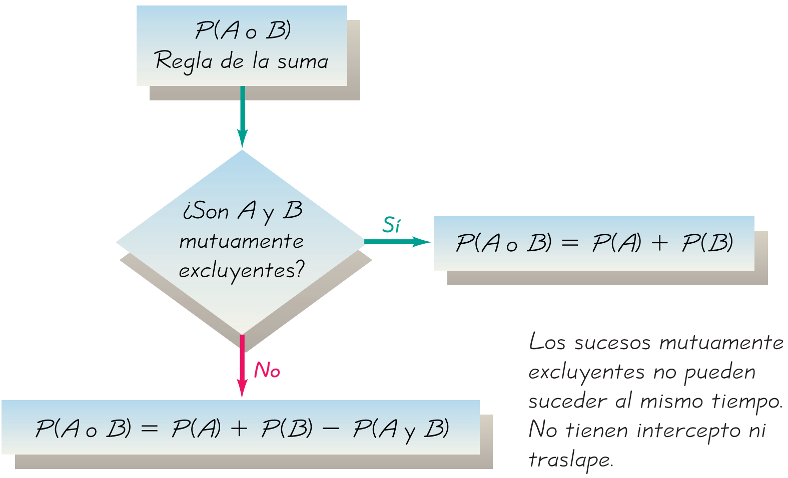


FIGURA 3-6 Aplicación de la regla de la suma

Definición

Los sucesos A y B son **desarticulados** (o **mutuamente excluyentes**) cuando ambos no pueden ocurrir juntos.

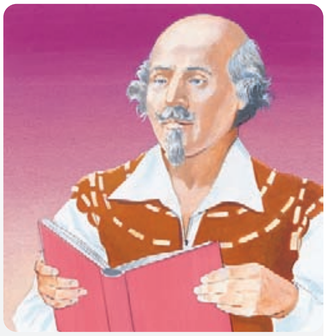
El diagrama de flujo de la figura 3-6 muestra cómo los sucesos mutuamente excluyentes afectan la regla de la suma.

EJEMPLO Ensayos clínicos de prueba de embarazo Remítase a la tabla 3-1, que se reproduce aquí para su conveniencia. Suponiendo que de entre 99 mujeres incluidas en el estudio, se selecciona una al azar, aplique la regla de la suma para calcular la probabilidad de seleccionar a una mujer que está embarazada o que tuvo un resultado de prueba positivo.

Tabla 3-1 Resultados de prueba de embarazo

	Resultado de prueba positivo (indicó embarazo)	Resultado de prueba negativo (no indicó embarazo)
La mujer está embarazada	80	5
La mujer no está embarazada	3	11

SOLUCIÓN Revisando la tabla, fácilmente advertimos que hay 88 mujeres embarazadas o que marcaron positivo. Obtenemos este total de 88 sumando las mujeres embarazadas a las mujeres que marcaron positivo, teniendo cuidado de contar a las 80 mujeres embarazadas que marcaron positivo sólo una vez. Sería erróneo sumar las 85 mujeres embarazadas con las 83 mujeres que marcaron positivo, ya que el total de 168 incluiría a algunas mujeres dos veces; hay que considerar que son individuos que deben contarse sólo una vez. Dividiendo el total correcto de 88 entre el total general de 99, obtenemos este resultado: $P(\text{embarazada o positiva}) = 88/99 = 8/9$ o 0.889.



El vocabulario de Shakespeare

Según Bradley Efron y Ronald Thisted, los escritos de Shakespeare incluyen 31,534 palabras diferentes. Ellos usaron la teoría de la probabilidad para concluir que Shakespeare probablemente conocía al menos otras 35,000 palabras que no usó en sus escritos. Estimar el tamaño de una población es un problema importante que se encuentra con frecuencia en estudios ecológicos, pero el resultado que se presenta aquí es otra aplicación interesante. (Véase "Estimating the Number of Unseen Species: How Many Words Did Shakespeare Know?" en *Biometrika*, vol. 63, núm. 3).

Hay varias estrategias que usted podría usar en el ejemplo anterior, para contar a las mujeres que estaban embarazadas o que marcaron positivo. Cualquiera de los siguientes funcionaría:

- Resalte con color las casillas que representan mujeres que estaban embarazadas o que marcaron positivo; luego, sume los números en estas casillas, teniendo cuidado de sumar cada número sólo una vez. Este método nos da

$$3 + 80 + 5 = 88$$

- Sume las 85 mujeres embarazadas a las 83 mujeres que marcaron positivo, pero compense el doble conteo restando las 80 mujeres embarazadas que marcaron positivo. Este método da un resultado de

$$85 + 83 - 80 = 88$$

- Comience con el total de 85 mujeres embarazadas; después, sume aquellas que marcaron positivo y que no se incluyeron aún en ese total, para obtener un resultado de

$$85 + 3 = 88$$

Estudie cuidadosamente el ejemplo anterior, que clarifica esta característica esencial de la regla de la suma: "o" implica suma, y la suma debe hacerse sin doble conteo.

Los puntos clave de esta sección se resumen como sigue:

1. Para calcular $P(A \text{ o } B)$, empiece por asociar "o" con la suma.
2. Considere si los sucesos A y B son mutuamente excluyentes, es decir, ¿pueden ocurrir al mismo tiempo? Si no son mutuamente excluyentes (es decir, pueden ocurrir al mismo tiempo), asegúrese de evitar (o por lo menos compensar) el doble conteo cuando sume las probabilidades relevantes. Entendiendo la importancia de no contar por duplicado cuando calcule $P(A \text{ o } B)$, no será necesario tener que determinar el valor de $P(A) + P(B) - P(A \text{ y } B)$.

Los errores que se cometen al aplicar la regla de la suma a menudo implican doble conteo, es decir, sucesos que no son mutuamente excluyentes se tratan como si lo fueran. Una indicación de semejante error es una probabilidad total mayor que 1; sin embargo, los errores que se relacionan con la regla de la suma no siempre hacen que la probabilidad total rebase 1.

Sucesos complementarios

En la sección 3-2, definimos el complemento del suceso A y lo denotamos como $\overline{A}$. Decimos que consiste en todos los resultados donde el suceso A no puede ocurrir. Los sucesos A y $\overline{A}$ son por implicación mutuamente excluyentes, ya que es imposible que un suceso y su complemento ocurran al mismo tiempo. Además, es posible estar absolutamente seguros de que A ocurre, o bien, que no ocurre, lo que implica que debe ocurrir A o $\overline{A}$. Estas observaciones nos permiten aplicar la regla de la suma para sucesos mutuamente excluyentes, así que:

$$P(A \text{ o } \overline{A}) = P(A) + P(\overline{A}) = 1$$

Justificamos $P(A \text{ o } \overline{A}) = P(A) + P(\overline{A})$ señalando que A y $\overline{A}$ son mutuamente excluyentes; justificamos el total de 1 por nuestra certeza absoluta de que A ocurre, o bien, no ocurre. Este resultado de la regla de la suma da lugar a las siguientes tres expresiones equivalentes.

Regla de los sucesos complementarios

$$P(A) + P(\overline{A}) = 1$$
$$P(\overline{A}) = 1 - P(A)$$
$$P(A) = 1 - P(\overline{A})$$

La figura 3-7 es una representación visual de la relación entre $P(A)$ y $P(\overline{A})$.

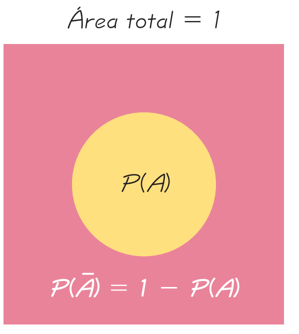


FIGURA 3-7 Diagrama de Venn para el complemento del suceso *A*

EJEMPLO En la realidad, cuando nace un bebé, $P(\text{niño}) = 0.5121$. Calcule $P(\overline{\text{niño}})$.

SOLUCIÓN Usando la regla de los sucesos complementarios, tenemos

$$P(\overline{\text{niño}}) = 1 - P(\text{niño}) = 1 - 0.5121 = 0.4879.$$

Es decir, la probabilidad de no tener un niño, que es la misma que la de tener una niña, es de 0.4879.

La principal ventaja de la regla de los sucesos complementarios es que permite simplificar ciertos problemas considerablemente. Ilustraremos esta ventaja en la sección 3-5.

3-3 Destrezas y conceptos básicos

Determine si los sucesos son mutuamente excluyentes. *Para cada inciso de los ejercicios 1 y 2, ¿son los dos eventos mutuamente excluyentes en un mismo experimento?* (Sugerencia: *Considere "mutuamente excluyentes" como equivalente a "separados" o "sin traslape".*)

1. a. Seleccionar aleatoriamente a un cirujano cardiaco.
Seleccionar aleatoriamente a un médico de sexo femenino.
b. Seleccionar aleatoriamente a una estudiante universitaria.
Seleccionar aleatoriamente a un estudiante universitario que conduzca motocicleta.
c. Seleccionar aleatoriamente a una persona que se trata con el fármaco para reducir el colesterol Lipitor.
Seleccionar aleatoriamente a una persona de un grupo control que no recibió tratamiento.

2. a. Seleccionar aleatoriamente a un padre de familia que esta noche, a las 8:15, se encuentre viendo la cadena NBC en la televisión.
Seleccionar aleatoriamente a un padre de familia que esta noche, a las 8:15, se encuentre viendo la cadena CBS en la televisión.
b. Recibir una llamada telefónica en respuesta a una encuesta voluntaria, de una persona que se opone a todos los impuestos gubernamentales.
Recibir una llamada telefónica en respuesta a una encuesta voluntaria, de una persona que aprueba todos los impuestos gubernamentales.
c. Seleccionar aleatoriamente a un senador de Estados Unidos que ocupe actualmente un puesto en el gobierno.
Seleccionar aleatoriamente a una mujer funcionaria electa.

3. Cálculo de complementos
a. Si $P(A) = 0.05$, calcule $P(\overline{A})$
b. Según datos de la oficina de censos de Estados Unidos, cuando se selecciona al azar a una mujer de 25 años de edad, hay una probabilidad de 0.218 de que tenga

una licenciatura. Si se selecciona al azar a una mujer de 25 años, calcule la probabilidad de que no tenga licenciatura.

4. Cálculo de complementos
a. Calcule $P(\overline{A})$ puesto que $P(A) = 0.0175$.
b. Una encuesta de Reuters/Zogby mostró que el 61% de los estadounidenses creen que existe vida en otro sitio en la galaxia. ¿Cuál es la probabilidad de seleccionar al azar a algún estadounidense que no crea esto?

5. Uso de la regla de la suma Remítase a la figura 3-3 y calcule la probabilidad de que al seleccionar aleatoriamente uno de los chícharos, obtenga uno con vaina verde o flor blanca.

6. Uso de la regla de la suma Remítase a la figura 3-3 y calcule la probabilidad de que al seleccionar aleatoriamente uno de los chícharos, obtenga uno con vaina amarilla o flor morada.

7. Día nacional de la estadística Si se selecciona una persona al azar, calcule la probabilidad de que su cumpleaños no sea el 18 de octubre, que es el Día Nacional de la Estadística en Japón. Ignore los años bisiestos.

8. Cumpleaños y complemento Si selecciona a una persona al azar, calcule la probabilidad de que su cumpleaños no sea en octubre. Ignore años bisiestos.

En los ejercicios 9 a 12 utilice los datos de la siguiente tabla, que resume resultados del hundimiento del Titanic.

	Hombres	Mujeres	Niños	Niñas
Sobrevivientes	332	318	29	27
Muertos	1360	104	35	18

9. Pasajeros del *Titanic* Si selecciona al azar a uno de los pasajeros del *Titanic*, calcule la probabilidad de que sea una mujer o una niña.

10. Pasajeros del *Titanic* Si selecciona al azar a uno de los pasajeros del *Titanic*, calcule la probabilidad de que sea un hombre o una persona que sobrevivió al hundimiento.

11. Pasajeros del *Titanic* Si selecciona al azar a uno de los pasajeros del *Titanic*, calcule la probabilidad de que sea un niño o un sobreviviente.

12. Pasajeros del *Titanic* Si selecciona al azar a uno de los pasajeros del Titanic, calcule la probabilidad de que sea una mujer o alguna persona que no sobrevivió al hundimiento.

Uso de la regla de la suma con grupos sanguíneos. *En los ejercicios 13 a 20 remítase a la gráfica adjunta, que describe los grupos sanguíneos y los tipos de Rh de 100 personas (según datos del Greater New York Blood Program). En cada caso, suponga que se selecciona uno de los 100 sujetos aleatoriamente; calcule la probabilidad que se indica.*

13. P(no grupo A)

14. P(tipo Rh^-)

15. P(grupo A o tipo Rh^-)

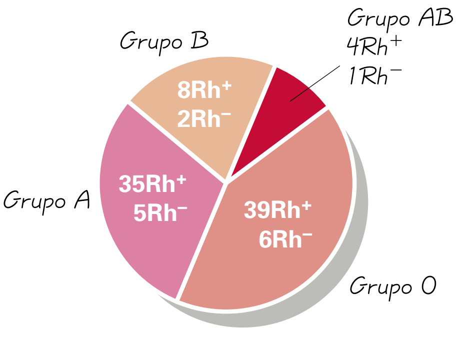
16. P(grupo A o grupo B)

17. P(no tipo Rh^+)

18. P(grupo B o tipo Rh^+)

19. P(grupo AB o tipo Rh^+)

20. P(grupo A u O o tipo Rh^+)



21. Datos de automóviles Remítase al conjunto de datos 22 del Apéndice B. Si uno de los 20 automóviles se selecciona al azar, calcule la probabilidad de que tenga transmisión manual o seis cilindros.

22. Tabaquismo y género Remítase al conjunto de datos 4 del Apéndice B. Si uno de los 107 sujetos de estudio se selecciona al azar, calcule la probabilidad de tener un varón o un fumador.

23. Resistencia a la encuesta Las empresas que realizan encuestas se interesan en los niveles decrecientes de cooperación de las personas que se contactan para que las encuesten. Un encuestador contacta a 84 individuos de entre 18 y 21 años, y descubre que 73 responden y 11 se rehúsan a hacerlo. Cuando se contacta a 275 personas de entre 22 y 29 años, 255 responden y 20 se rehúsan (según datos de "I Hear You Knocking but You Can't Come In", de Fitzgerald y Fuller, *Sociological Methods and Research*, vol. 11, núm. 1). Suponga que se selecciona al azar a 1 de las 359 personas. Calcule la probabilidad de que sea una persona en el rango de edad de 18 a 21 años o alguien que rechaza responder.

24. Resistencia a la encuesta Remítase al mismo conjunto de datos que se utilizó en el ejercicio 23. Suponga que se selecciona al azar a 1 de las 359 personas, calcule la probabilidad de que sea una persona en el rango de 18 a 21 años o que sí respondió.

3-3 Más allá de lo básico

25. Determine si los sucesos son mutuamente excluyentes

a. Si $P(A) = 3/11$, $P(B) = 4/11$ y $P(A \text{ o } B) = 7/11$, ¿qué puede inferir acerca de los sucesos A y B?

a. Si $P(A) = 5/18$, $P(B) = 11/18$, y $P(A \text{ o } B) = 13/18$, ¿qué puede inferir acerca de los sucesos A y B?

26. Sucesos mutuamente excluyentes Si los sucesos A y B son mutuamente excluyentes y los sucesos B y C también son mutuamente excluyentes, ¿tienen que ser mutuamente excluyentes los sucesos A y C? Dé un ejemplo que fundamente su respuesta.

27. *O* exclusivo ¿Como se transformaría la regla de la suma si el *o exclusivo* se usara en lugar del *o inclusivo*? En esta sección se explicó que el *o exclusivo* significa uno u otro, pero no ambos.

28. Extensión de la regla de la suma La regla formal de la suma, que se incluye en esta sección, expresa la probabilidad de A o B como sigue: $P(A \text{ o } B) = P(A) + P(B) - P(A \text{ y } B)$. Extienda esta regla formal para desarrollar una expresión aplicable a $P(A \text{ o } B \text{ o } C)$. (*Sugerencia:* Dibuje un diagrama de Venn).

3-4 Regla de la multiplicación: fundamentos

En la sección 3-3 presentamos la regla de la suma para calcular $P(A \text{ o } B)$, es decir, la probabilidad de que un ensayo tenga un resultado de A o B o ambos. El objetivo de esta sección es desarrollar una regla para calcular $P(A \text{ y } B)$, esto es, la probabilidad de que el suceso A ocurra en un primer ensayo y el suceso B ocurra en un segundo ensayo.

Sentenciados por probabilidad

Un testigo describió a una asaltante de Los Ángeles como una mujer de raza caucásica con pelo rubio, peinado en cola de caballo, que escapó en un automóvil amarillo que conducía un hombre afroamericano que usaba barba y bigote. Janet y Malcom Collins se ajustaban a tal descripción y se les condenó con fundamento en el testimonio de que hay aproximadamente una posibilidad en 12 millones de que cualquier pareja tenga tales características. Se estimó que la probabilidad de poseer un automóvil amarillo es de 1/10, en tanto que las demás probabilidades se estimaron en 1/4, 1/10, 1/3, 1/10 y 1/1,000. Más tarde, las condenas se anularon, cuando se señaló que no se presentó evidencia que apoyara las probabilidades que se estimaron o la independencia de los sucesos. Sin embargo, puesto que la pareja no se seleccionó aleatoriamente, se cometió un error grave al no considerar la probabilidad de que hubiera *otras* parejas en la misma región con las mismas características.

Notación

$P(A \text{ y } B) = P(\text{el suceso } A \text{ ocurre en un primer ensayo y el suceso } B \text{ ocurre en un segundo ensayo})$

En la sección 3-3 asociamos *o* con sumar; en esta sección asociaremos *y* con multiplicar. Veremos que $P(A \text{ y } B)$ implica la multiplicación de probabilidades, y que en ocasiones deberemos ajustar la probabilidad del suceso B para discernir el resultado del suceso A.

La teoría de la probabilidad se utiliza extensamente en el análisis y el diseño de pruebas estandarizadas, como SAT, ACT, LSAT (para leyes) y MCAT (para medicina). Para facilitar la calificación, las pruebas de este tipo suelen incluir preguntas del tipo verdadero/falso o de opción múltiple. Supongamos que el primer reactivo de un examen es del tipo verdadero/falso, y que el segundo es de opción múltiple con cinco respuestas posibles (*a*, *b*, *c*, *d*, *e*). Usemos los dos reactivos siguientes. ¡Intente responderlos!

1. Verdadero o falso: Una libra de plumas pesa más que una libra de oro.
2. Entre lo siguiente, ¿qué es lo que tiene la mayor influencia en la sociedad moderna?
   - **a.** El control remoto
   - **b.** Este libro
   - **c.** Las computadoras
   - **d.** Los tenis con luces en el tacón
   - **e.** Las recepcionistas coquetas

Las respuestas a los dos reactivos son V (de "verdadero") y c. (La primera pregunta es verdadera. Los pesos de las plumas se expresan en libras avoirdupois, pero los pesos del oro se expresan en libras troy.) Calculemos la probabilidad de que si alguna persona hace suposiciones al azar para ambas respuestas, la respuesta al primer reactivo sea correcta y la respuesta al segundo reactivo sea también correcta. Una forma de calcular esta probabilidad es elaborar una lista del espacio muestral, como sigue:

V,a	V,b	V,c	V,d	V,e
F,a	F,b	F,c	F,d	F,e

Si las respuestas son conjeturas al azar, tenemos que los 10 posibles resultados son igualmente probables, entonces

$$P(\text{ambas correctas}) = P(T \text{ y } c) = \frac{1}{10} = 0.1$$

Ahora note que $P(V \text{ y } c) = 1/10$, $P(V) = 1/2$, y $P(c) = 1/5$; por lo tanto, vemos que

$$\frac{1}{10} = \frac{1}{2} \cdot \frac{1}{5}$$

de modo que

$$P(T \text{ y } c) = P(T) \times P(c)$$

Esto sugiere que, en términos generales, $P(A \text{ y } B) = P(A) \cdot P(B)$, pero antes de hacer dicha generalización, consideremos otro ejemplo.

Por lo pronto, notamos que los diagramas de árbol suelen utilizarse para determinar el número de resultados posibles en el espacio muestral. Un **diagrama de árbol** es una imagen gráfica de los resultados posibles de un procedimiento, que se muestran como líneas que emanan de un punto de partida. Estos diagramas son útiles para calcular el número de resultados posibles, cuando el número de posibilidades no es demasiado grande. El diagrama de árbol de la figura 3-8 muestra los resultados de los reactivos de verdadero/falso y opción múltiple. En la figura 3-8 vemos que si las dos respuestas son conjeturas al azar, las 10 ramas son igualmente probables, y la probabilidad de obtener el par correcto (V,c) es de 1/10. Para cada respuesta a la primera pregunta, hay cinco respuestas a la segunda. El número total de resultados es cinco dos veces, o sea, 10. El diagrama de árbol de la figura 3-8 ilustra la razón del uso de la multiplicación.

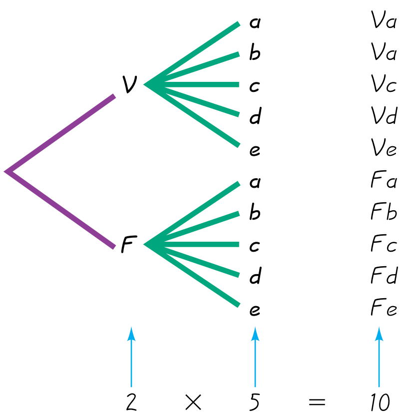


FIGURA 3-8 Diagrama de árbol de reactivos de examen

Nuestro primer ejemplo, el de los reactivos de verdadero/falso y opción múltiple, sugiere que $P(A \text{ y } B) = P(A) \cdot P(B)$; el siguiente ejemplo introducirá otro elemento importante.

EJEMPLO Experimento de genética En los famosos experimentos de hibridación de Mendel se emplearon chícharos, como los que se muestran en la figura 3-3, que se incluye en la sección 3-3 y se reproduce en la página siguiente. Si dos de los chícharos que se observan en la figura 3-3 se seleccionan al azar sin reemplazo, calcule la probabilidad de que la primera selección tenga una vaina verde y la segunda una vaina amarilla. (Es posible ignorar los colores de las flores en la parte superior).

SOLUCIÓN

Primera selección:	(porque hay 14 chícharos, ocho
$P(\text{vaina verde}) = 8/14$	de los cuales tienen vainas verdes)
Segunda selección:	(hay 13 chícharos sobrantes, seis
$P(\text{vaina amarilla}) = 6/13$	de los cuales tienen vainas amarillas)

continúa

Calificación perfecta en el sat

Si se selecciona al azar un sujeto que responde el SAT, ¿cuál es la probabilidad de elegir a una persona que obtenga una calificación perfecta? ¿Cuál es la probabilidad de obtener una calificación perfecta en el SAT, adivinando las respuestas? Se trata de dos preguntas muy diferentes.

En un año reciente, aproximadamente 1.3 millones de personas respondieron el SAT, y sólo 587 recibieron calificaciones perfectas de 1,600, entonces hay una probabilidad de 587 ÷ 1.3 millones, o alrededor de 0.000452, de seleccionar aleatoriamente a uno de los sujetos de la prueba y a una persona con una calificación perfecta. Sólo una parte del SAT incluye 35 preguntas de opción múltiple, y la probabilidad de responder a todas ellas correctamente adivinando es de $(1/5)^{35}$, cantidad tan pequeña que, cuando se escribe como un decimal, resultan 24 ceros después del punto decimal.

Con P(primer chícharo con vaina verde) $= 8/14$ y P(segundo chícharo con vaina amarilla) $= 6/13$, tenemos

$$P(\text{primer chícharo con vaina verde y segundo chícharo con vaina amarilla}) = \frac{8}{14} \cdot \frac{6}{13} \approx 0.264$$

El punto clave es que se tiene que ajustar la probabilidad del segundo suceso para reflejar el resultado del primer suceso. Ya que el segundo chícharo se selecciona sin reemplazar el primero, la segunda probabilidad debe tomar en cuenta el resultado de la primera selección de un chícharo con vaina verde. Después de que se ha seleccionado un chícharo con vaina verde en el primer ensayo, sólo quedan 13 chícharos y seis de ellos tienen vainas amarillas, entonces la segunda selección nos da: P(chícharo con vaina amarilla) $= 6/13$.

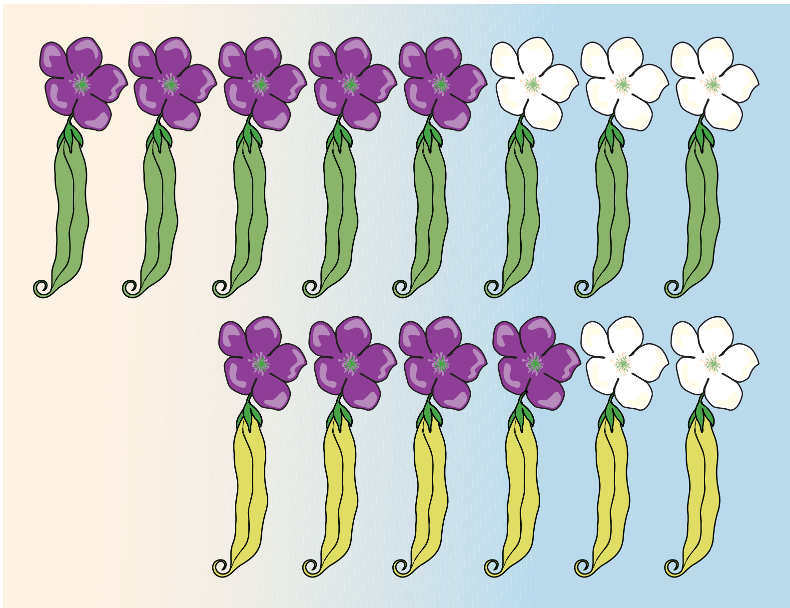


FIGURA 3-3 **Chícharos usados en un estudio de genética**

Este ejemplo manifiesta el importante principio de que *la probabilidad del segundo suceso* B *debe tomar en cuenta el hecho de que el primer suceso* A *ya ocurrió*. Este principio suele expresarse usando la notación siguiente.

Notación para la probabilidad condicional

$P(B \mid A)$ representa la probabilidad de que un suceso B ocurra después de admitir que el suceso A ya ocurrió. (Es posible leer $B \mid A$ como "B dado A").

Definiciones

Dos sucesos A y B son **independientes** cuando la ocurrencia de uno no afecta la probabilidad de la ocurrencia del otro. (De manera similar, algunos sucesos son independientes si la ocurrencia de cualquiera no afecta las probabilidades de la ocurrencia de los demás). Si A y B no son independientes, se dice que son **dependientes**.

LA ESTADÍSTICA EN LAS NOTICIAS

Recomendación para la lotería

Un columnista del diario *New York Daily News*, Stephen Allensworth, hace poco dio recomendaciones para seleccionar números en el juego *New York State's Daily Numbers*. En la descripción de un sistema para ganar, escribió que "comprende números dobles asociados con dígitos fríos. (Un dígito frío es uno que sale una vez o no sale nunca en un periodo de siete días)". Allensworth procedió a identificar algunos números específicos que "tienen una excelente probabilidad de salir esta semana".

Allensworth supone que algunos números están "rezagados", pero la selección de números en la lotería es independiente de los resultados pasados. El sistema que describe no tiene bases reales y no funcionará. Los lectores que siguen una recomendación tan pobre como ésta, se están dejando engañar y perderán más dinero, ya que creen erróneamente que sus probabilidades de ganar mejoran.

Por ejemplo, jugar a la lotería de California y después a la lotería de Nueva York son sucesos *independientes*, porque el resultado de la lotería de California no surte efecto alguno en las probabilidades de los resultados de la lotería de Nueva York. En contraste, el suceso de intentar arrancar su automóvil y el suceso de llegar a clase a tiempo son sucesos *dependientes*, porque el resultado del intento de arrancar su automóvil afecta la probabilidad de llegar a clase a tiempo.

Con la notación y las definiciones anteriores, junto con los principios ilustrados en los ejemplos, resumimos el concepto clave de la sección como la siguiente *regla formal de la multiplicación*, pero se recomienda que usted trabaje con la *regla intuitiva de la multiplicación*, que tiene más probabilidades de manifestar comprensión que el uso a ciegas de una fórmula.

Regla formal de la multiplicación

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

Si A y B son sucesos independientes, $P(B|A)$ es realmente lo mismo que $P(B)$. (Para hacer un estudio más amplio y determinar si los sucesos son independientes o dependientes, véase el apartado "Prueba de independencia", en la sección 3-5. Por lo pronto, trate de entender el concepto básico de independencia y la forma en que afecta las probabilidades calculadas). Observe la siguiente *regla intuitiva de la multiplicación*. (Véase también la figura 3-9).

Regla intuitiva de la multiplicación

Cuando se trata de calcular la probabilidad de que el suceso A ocurra en un ensayo y el suceso B ocurra en el siguiente ensayo, multiplique la probabilidad del suceso A por la probabilidad del suceso B, pero asegúrese de que la probabilidad del suceso B tome en cuenta la ocurrencia previa del suceso A.

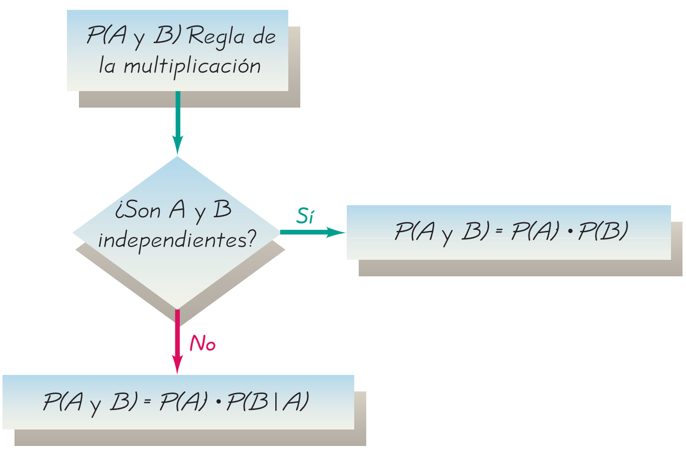


FIGURA 3-9 Aplicación de la regla de la multiplicación

Motores a reacción independientes

Poco después de salir de Miami, el vuelo 855 de Eastern Airlines tuvo que apagar un motor porque se encendió el indicador de baja presión de aceite. Cuando el jet L-1011 regresaba a Miami para aterrizar, los indicadores de baja presión de los otros dos motores también se encendieron. Entonces falló otro motor y luego falló el último motor que funcionaba. El jet descendió sin propulsión desde 13,000 pies hasta 4,000 pies, por lo que la tripulación logró arrancar un motor y la aeronave, con las 172 personas a bordo, que aterrizaron con seguridad. Con motores a reacción independientes, la probabilidad de que los tres fallen es de sólo 0.0001^3, es decir, alrededor de una en un billón. La FAA averiguó que el mismo mecánico que cambió el aceite de los tres motores se equivocó al reemplazar los anillos de sello del tapón de aceite. El empleo de un solo mecánico hizo que el funcionamiento de los motores se volviera dependiente, situación que se corrigió exigiendo que los motores reciban mantenimiento por mecánicos diferentes.

EJEMPLO Bienes dañados Telektronics fabrica computadoras, televisores, reproductores de CD y otros productos electrónicos. Cuando los artículos que se envían se dañan, las causas del daño se clasifican como agua (A), compresión (C), perforación (P) o marcas en la caja (M). Abajo se encuentra una lista de las causas codificadas de cinco artículos que se dañaron. Una analista de control de calidad quiere seleccionar aleatoriamente dos artículos para elaborar una investigación más amplia. Calcule la probabilidad de que el primer artículo fuese dañado por compresión (C) y el segundo también por lo mismo (C). Suponga que las selecciones se hacen *a*) con reemplazo; *b*) sin reemplazo.

A C C P M

SOLUCIÓN

a. Si los dos artículos se seleccionan con reemplazo, las dos selecciones son independientes, ya que al segundo suceso no le afecta el primer resultado. En cada una de las dos selecciones hay dos artículos que se dañaron por compresión (C) entre los cinco; entonces, tenemos

$$P(\text{el primer artículo es C y el segundo artículo es C}) = \frac{2}{5} \cdot \frac{2}{5} = \frac{4}{25} \text{ o } 0.16$$

b. Si los dos artículos se seleccionan sin reemplazo, las dos selecciones son dependientes porque el segundo suceso se afectó por el primer resultado. En la primera selección, a dos de los cinco artículos los daño la compresión (C). Después de seleccionar un artículo dañado por compresión, estamos dejando cuatro artículos incluyendo a uno al que dañó también la compresión. Por lo tanto, tenemos

$$P(\text{el primer artículo es C y el segundo artículo es C}) = \frac{2}{5} \cdot \frac{1}{4} = \frac{2}{20} = \frac{1}{10} \text{ o } 0.1$$

Nótese que en este caso ajustamos la segunda probabilidad para tomar en cuenta la selección de un artículo al que dañó la compresión (C) en el primer resultado. Después de seleccionar C la primera vez, había sólo un C entre los cuatro artículos que quedaban.

Hasta aquí ya analizamos dos sucesos, pero la regla de la multiplicación puede extenderse fácilmente a varios sucesos. En general, la probabilidad de cualquier secuencia de sucesos independientes es simplemente el producto de sus probabilidades correspondientes. Por ejemplo, la probabilidad de lanzar una moneda tres veces y obtener siempre caras es de $0.5 \cdot 0.5 \cdot 0.5 = 0.125$. También es posible extender la regla de la multiplicación para aplicarla a varios sucesos dependientes; simplemente hay que ajustar las probabilidades conforme se avanza. Por ejemplo, la probabilidad de sacar cuatro cartas diferentes (sin reemplazo) de un mazo revuelto y que todas sean ases es

$$\frac{4}{52} \cdot \frac{3}{51} \cdot \frac{2}{50} \cdot \frac{1}{49} = 0.00000369$$

El inciso *b* del último ejemplo implicó la selección de artículos sin reemplazo; por lo tanto, tratamos los sucesos como dependientes. Sin embargo, es común tratar sucesos como independientes cuando se toman *muestras pequeñas* de *poblaciones grandes*. En tales casos, es raro seleccionar el mismo elemento dos veces. He aquí un lineamiento común:

Si el tamaño de la muestra no es mayor que el 5% del tamaño de la población, trate las selecciones como si fueran *independientes* (si las selecciones se hacen sin reemplazo, de modo que sean técnicamente dependientes).

Los encuestadores usan este lineamiento cuando encuestan apenas a 1,000 adultos de poblaciones de millones. Ellos suponen independencia, aunque toman la muestra sin reemplazo. El siguiente ejemplo es otra ilustración de dicho lineamiento. También ejemplifica cómo se utiliza la probabilidad para probar una afirmación hecha acerca de una población. Nos da idea del importante procedimiento de *prueba de hipótesis* que se estudiará en el capítulo 7.

EJEMPLO Control de calidad Una gerente de producción de Telektronics afirma que su nuevo proceso de manufactura de DVD es mejor porque la tasa de defectos es menor del 3%, que fue la tasa de defectos en el pasado. Para fundamentar su afirmación, ella fabrica un lote de 5,000 DVD; después, selecciona al azar 200 de ellos para probarlos, con el resultado de que no hay defectos en ninguno de los 200 DVD que se seleccionaron. Suponiendo que el nuevo método tuviera la misma tasa de defectos del 3% como en el pasado, calcule la probabilidad de que no haya defectos en los 200 DVD. Con base en el resultado, ¿hay suficiente evidencia para fundamentar la afirmación de la gerente de que su nuevo proceso es mejor?

SOLUCIÓN La probabilidad de que no tengan defectos es la misma que la probabilidad de que los 200 DVD estén en buen estado. Por lo tanto, queremos encontrar *P*(todos los 200 DVD en buen estado). También suponemos que la tasa de defectos es del 3% para observar si el resultado de cero defectos, por su probabilidad, llegaría a ocurrir fácilmente y poder así compararlo con el antiguo proceso de fabricación. Si la tasa de defectos es del 3%, tenemos $P(\text{DVD en buen estado}) = 0.97$. Los DVD que se seleccionaron fueron escogidos sin reemplazo, pero la muestra de 200 DVD es menor al 5% de la población de 5,000. Entonces, trataremos los sucesos como si fueran independientes. Obtenemos este resultado:

$$P(\text{1o en buen estado y 2o en buen estado y 3o en buen estado} \ldots \text{y 200o en buen estado})$$
$$= P(\text{DVD en buen estado}) \cdot P(\text{DVD buen estado}) \cdot \ldots \cdot P(\text{DVD en buen estado})$$
$$= 0.97 \cdot 0.97 \cdot \ldots \cdot 0.97$$
$$= 0.97^{200} = 0.00226$$

La baja probabilidad de 0.00226 indica que en lugar de obtener un resultado poco común, con una tasa de defecto del 3%, una explicación más razonable es que no ocurrieron defectos porque la tasa de defectos es realmente menor que el 3%. Debido a que hay una probabilidad tan pequeña (0.00226) de producir todos los DVD en buen estado, con un tamaño de muestra de 200 y una tasa de defectos del 3%, tenemos suficiente evidencia para concluir que el nuevo método es mejor.

Los fundamentos de las reglas de la suma y de la multiplicación se resumen como sigue:

- En la regla de la suma, la palabra "o" en *P*(*A* o *B*) sugiere una suma. Sume *P*(*A*) y *P*(*B*), siendo cuidadoso para hacerlo de forma que cada resultado se cuente sólo una vez.

LA ESTADÍSTICA EN LAS NOTICIAS

Redundancia

Es posible mejorar considerablemente la confiabilidad de los sistemas con la redundancia de componentes críticos. Los automóviles de carreras de las series de la NASCAR Winston Cup tienen dos sistemas de ignición para que, si uno falla, haya otro de reserva. Los aviones poseen dos sistemas eléctricos independientes, y los que se usan para vuelos instrumentales suelen tener dos radios distintos. La siguiente cita se tomó de un artículo de *Popular Science* sobre los aviones antirradar: "Un avión construido en buena parte con fibra de carbono fue el Lear Fan 2100, que debía llevar dos transpondedores de radar. La razón es que si fallaba una unidad de transpondedor, el avión seguiría siendo casi invisible para el radar". Tal redundancia es una aplicación de la regla de la multiplicación de la teoría de la probabilidad. Si un componente tiene una probabilidad de 0.001 de fallar, la probabilidad de que dos componentes independientes fallen es de sólo 0.000001.

- En la regla de la multiplicación, la palabra "y" en $P(A \text{ y } B)$ sugiere una multiplicación. Multiplique $P(A)$ y $P(B)$, pero asegúrese de que la probabilidad del suceso B toma en cuenta la ocurrencia previa del suceso A.

3-4 Destrezas y conceptos básicos

Identificación de sucesos como independientes o dependientes*. En los ejercicios 1 y 2 clasifique cada par de sucesos como independientes o dependientes.*

1. **a.** Tirar un dado y obtener un 5.
Lanzar una moneda y obtener cara.
b. Seleccionar aleatoriamente a un televidente que ve *Monday Night Football.*
Seleccionar aleatoriamente a un segundo televidente que ve *Monday Night Football.*
c. Usar pantalones cortos a cuadros con calcetines negros y sandalias.
Pedir a alguien una cita y tener una respuesta positiva.

2. **a.** Descubrir que su calculadora no funciona.
Descubrir que su refrigerador no funciona.
b. Descubrir que la luz de su cocina no funciona.
Descubrir que su refrigerador no funciona.
c. Beber hasta deteriorar su capacidad de conducir.
Verse involucrado en un accidente automovilístico.

3. Moneda y dado Calcule la probabilidad de que al lanzar una moneda y tirar un dado, los resultados sean cruz y 3.

4. Letra y dígito La propietaria de una computadora nueva crea una contraseña que consta de dos caracteres. Ella selecciona al azar una letra del alfabeto para el primer carácter y un dígito (0, 1, 2, 3, 4, 5, 6, 7, 8, 9) para el segundo. ¿Cuál es la probabilidad de que su contraseña sea "K9"? ¿Sería eficaz esta contraseña como obstáculo contra alguien que trate de tener acceso a su computadora?

5. Aplicación de la regla de la multiplicación Si dos de los elementos que se muestran abajo se seleccionan al azar, calcule la probabilidad de que ambos elementos sean de color verde. Estos elementos se utilizan en pruebas de percepción.

rojo amarillo verde rojo azul amarillo

a. Suponga que el primer elemento se reemplaza antes de seleccionar el segundo.
b. Suponga que el primer elemento no se reemplaza antes de seleccionar el segundo.

6. Aplicación de la regla de la multiplicación Usando los mismos seis elementos del ejercicio 5, calcule la probabilidad de seleccionar al azar tres elementos y obtener uno de color rojo en la primera selección, uno de color verde en la segunda y un elemento azul en la tercera.
a. Suponga que cada elemento se reemplaza antes de que se seleccione el siguiente.
b. Suponga que ninguno de los elementos que se seleccionaron se reemplaza antes de que los otros sean seleccionados.

7. Máscaras antigás defectuosas La revista *Time* reportó que cuando se probaron 19,218 máscaras antigás en divisiones de la milicia de Estados Unidos, se encontró que 10,322 estaban defectuosas (según datos de la Organización Mundial de la Salud). Si una investigación más a fondo comienza por la selección aleatoria de dos máscaras antigás de esta población, calcule la probabilidad de que ambas estén defectuosas.

a. Suponga que la primera máscara antigás se reemplaza antes de seleccionar la segunda.

b. Suponga que la primera máscara antigás no se reemplaza antes de seleccionar la segunda.

c. Compare los resultados que se obtuvieron en *a*) y *b*).

d. Dada la alternativa entre seleccionar con reemplazo y sin reemplazo, ¿cual alternativa es más lógica para tal situación? ¿Por qué?

8. Uso de ropa naranja de cazador Un estudio de heridas de caza en relación con el uso de ropa naranja "de cazador" mostró que de 123 cazadores heridos al confundirlos con presas, seis usaban ropa naranja (según datos de los Centers for Disease Control). Si un estudio de seguimiento comenzara con la selección aleatoria de cazadores de esta muestra de 123, calcule la probabilidad de que los primeros dos cazadores que se seleccionaron usaran ropa naranja.

a. Suponga que el primer cazador se reemplaza antes de que el siguiente se seleccione.

b. Suponga que el primer cazador no se reemplaza antes de que el segundo cazador se seleccione.

c. Dada la alternativa entre seleccionar con reemplazo y sin reemplazo, ¿cuál es más lógica para esta situación? ¿Por qué?

9. Probabilidad y adivinar Una profesora de psicología hace un examen sorpresa que consta en 10 reactivos verdadero/falso; establece que para aprobar se requieren al menos siete respuestas correctas. Suponga que un estudiante que no se preparó adopta la cuestionable estrategia de adivinar cada respuesta.

a. Calcule la probabilidad de que las primeras siete respuestas sean correctas y las últimas tres sean incorrectas.

b. ¿La probabilidad del inciso *a*) es igual a la probabilidad de aprobar? ¿Por qué?

10. Selección de senadores en Estados Unidos En el 107o Congreso, el Senado consta de 13 mujeres y 87 hombres. Si un cabildero de la industria del tabaco selecciona al azar a tres diferentes senadores, ¿cuál es la probabilidad de que sean mujeres? ¿Sería probable que un cabildero usara la selección aleatoria en esta situación?

11. Cumpleaños coincidentes

a. El autor nació el 27 de noviembre. ¿Cuál es la probabilidad de que otras dos personas que se seleccionen al azar nacieran también el 27 de noviembre? (Ignore los años bisiestos).

b. ¿Cuál es la probabilidad de que dos personas que se seleccionaron al azar tengan la misma fecha de cumpleaños? (Ignore los años bisiestos).

12. Cumpleaños coincidentes

a. Una pareja atrajo la atención de los medios de comunicación, ya que sus tres hijos, que nacieron en años diferentes, llegaron al mundo el 4 de julio. Ignorando los años bisiestos, calcule la probabilidad de que tres personas seleccionadas al azar nacieran el 4 de julio. ¿Es la probabilidad lo suficientemente baja como para que un suceso como éste no tenga probabilidades de ocurrir, en algún lugar de Estados Unidos, en el transcurso de varios años?

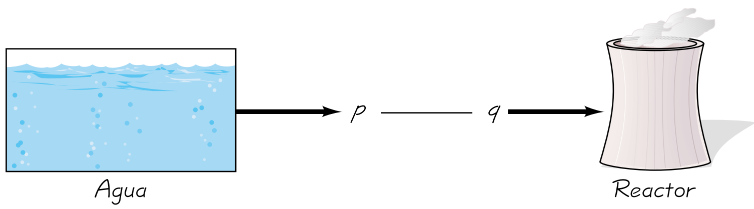
b. Ignore los años bisiestos y calcule la probabilidad de que tres personas que se seleccionen al azar tengan todas la misma fecha de cumpleaños.

13. Muestreo de aceptación Con cierto método de un procedimiento que se llama *muestreo de aceptación*, se selecciona aleatoriamente y sin reemplazo una muestra de artículos, el lote completo se acepta si cada artículo en la muestra es aprobado. La Niko Elctronics Company acaba de fabricar 5,000 CD, de los cuales el 3% están defectuosos. Si se seleccionan al azar 12 de estos CD para probarlos, ¿cuál es la probabilidad de que se acepte el lote completo?

14. Nivel de confianza de una encuesta En las encuestas de opinión pública es común manejar un "nivel de confianza" del 95%, lo que quiere decir que hay un 0.95 de probabilidad de que los resultados de la encuesta sean precisos dentro de los márgenes de error que se consideró. Si cinco organizaciones diferentes realizan encuestas independientes, ¿cuál es la probabilidad de que las cinco sean precisas dentro de los márgenes de error que se consideraron? ¿Sugiere el resultado que con un nivel de confianza del 95% es posible esperar que casi todas las encuestas estén dentro del margen de error que se consideró?

15. Prueba de efectividad de un método de selección de género Descubrimientos recientes parecen hacer posible que las parejas incrementen, de forma impresionante, la posibilidad de tener un hijo con el género de su elección. En una prueba de un método de selección del género, 10 parejas desean tener niñas. Si este método de selección del género no tuviera efecto, ¿cuál es la probabilidad de que 10 bebés sean todos niñas? Si en realidad resultan 10 niñas en 10 hijos, ¿parece ser efectivo este método de selección de género? ¿Por qué?

16. Confiabilidad de un reactor nuclear En un reactor nuclear se utilizan sensores remotos para controlar cada una de dos válvulas separadas e independientes, que se abren para abastecer agua para enfriamiento en caso de emergencia, las cuales se denotan por p y q. Cada válvula tiene un 0.9968 de probabilidad de abrirse cuando se le dispara. Para la configuración dada, calcule la probabilidad de que cuando ambos sensores se disparen, el agua fluya a través del sistema y ocurra enfriamiento. ¿El resultado es suficientemente alto para considerarse seguro?



17. La excusa de la llanta que se reventó Cuatro estudiantes que perdieron un examen ofrecen una excusa clásica, afirman que se le reventó una llanta al automóvil en el que los cuatro viajaban. En la reposición del examen, el maestro pide a cada uno de los estudiantes que identifique la llanta en particular que se reventó. Si ellos en realidad no tuvieron ninguna avería en los neumáticos, pero seleccionan al azar una llanta que supuestamente se reventó, ¿cuál es la probabilidad de que todos ellos escojan la misma llanta?

18. Identificación de la voz de un criminal En un caso legal en Riverhead, Nueva York, nueve víctimas de un crimen escucharon grabaciones de la voz de cinco hombres diferentes. Las nueve víctimas identificaron la misma voz como la del criminal. Si las identificaciones de voz se hubiesen hecho al azar, calcule la probabilidad de que las nueve víctimas seleccionaran a la misma persona. ¿Constituye esto una duda razonable?

19. Control de calidad Una gerente de producción de Telektronics afirma que su nuevo proceso de fabricación de reproductores de CD es mejor porque su tasa de defectos es más baja que el 2%, la tasa de defectos en el pasado. Para fundamentar su afirmación, ella fabrica un lote de 5,000 reproductores de CD, luego selecciona aleatoriamente 15 de ellos para probarlos, con el resultado de que no hay defectos en los 15 reproductores de CD que se seleccionaron. Con base en el resultado, ¿hay suficiente evidencia para fundamentar la afirmación de la gerente de que su nuevo proceso es mejor?

20. Redundancia El principio de la redundancia se utiliza cuando la confiabilidad de un sistema se mejora por medio de componentes redundantes o de respaldo. Suponga que su reloj despertador tiene un 0.975 de probabilidad de funcionar en cualquier mañana dada.

continúa

a. ¿Cuál es la probabilidad de que su reloj despertador no funcione en la mañana de un examen final importante?
b. Si usted tiene dos relojes despertadores como el descrito, ¿cuál es la probabilidad de que ambos fallen en la mañana de un examen final importante?
c. Con un reloj despertador, hay un 0.975 de probabilidad de ser despertados. ¿Cuál es la probabilidad de ser despertado si estamos usando dos relojes despertadores?

Resultados de prueba de embarazo *En los ejercicios 21 a 24 utilice los datos de la tabla 3-1, que se reproduce aquí.*

Tabla 3-1 Resultados de prueba de embarazo

	Resultado de prueba positivo (indicó embarazo)	Resultado de prueba negativo (no indicó embarazo)
La mujer está embarazada	80	5
La mujer no está embarazada	3	11

21. Resultado de prueba positivo Si se seleccionan al azar dos mujeres, calcule la probabilidad de que ambas pruebas den resultado positivo.

22. Embarazo Si se selecciona al azar una de las mujeres, calcule la probabilidad de elegir una que probó negativo o una que no está embarazada.

23. Embarazo Si se seleccionan al azar dos diferentes mujeres, calcule la probabilidad de que ambas estén embarazadas.

24. Resultado de prueba negativo Si se selecciona al azar tres mujeres, calcule la probabilidad de que todas obtuvieran un resultado negativo.

3-4 Más allá de lo básico

25. Las mismas fechas de cumpleaños Calcule la probabilidad de que dos personas no tengan la misma fecha de cumpleaños, cuando el número de personas que se seleccionó al azar es

a. 3 **b.** 5 **c.** 25

26. Género de hijos Si una pareja planea tener ocho hijos, calcule la probabilidad de que todos sean del mismo género.

27. Selección de cartas Se van a seleccionar dos cartas de un mazo revuelto, al azar y sin reemplazo. Calcule la probabilidad de obtener un as en la primera carta y una espada en la segunda carta.

28. Complementos y la regla de la suma
a. Desarrolle una fórmula para la probabilidad de no obtener A o B, ni ninguno de los dos, en un mismo ensayo. Esto es, calcule una expresión para $P(\overline{A \text{ o } B})$.
b. Desarrolle una fórmula para la probabilidad de no obtener A o no obtener B en un mismo ensayo. Esto es, calcule una expresión para $P(\overline{A} \text{ o } \overline{B})$.
c. Compare los resultados de los incisos *a* y *b*. ¿Es $P(\overline{A \text{ o } B}) = P(\overline{A} \text{ o } \overline{B})$?

Muestreo compuesto

En una ocasión, el ejército estadounidense hizo pruebas para determinar la presencia de sífilis en cada recluta, tomando una muestra de sangre individual que se analizaba por separado. Un investigador sugirió mezclar muestras de sangre por pares. Después de analizar los pares que se mezclaron, los reclutas con sífilis se identificaban volviendo a analizar las muestras de sangre individuales de los pocos pares que dieron resultados positivos en el análisis. El número total de análisis se redujo pareando especímenes de sangre, así que ¿por qué no colocarlos en grupos de tres o cuatro o más? Se usó la teoría de probabilidad para determinar el tamaño de grupo más eficiente y se desarrolló una teoría general para detectar los defectos en cualquier población. Dicha técnica se conoce como *muestreo compuesto*.

3-5 Regla de la multiplicación: complementos y probabilidad condicional

La sección 3-4 introdujo el concepto básico de la regla de la multiplicación; en esta sección extenderemos el uso de tal regla a dos aplicaciones especiales. Comencemos con situaciones en las cuales queremos identificar la probabilidad de que, entre varios ensayos, *uno al menos* dé un resultado que se especifica.

Complementos: La probabilidad de "uno al menos"

La regla de la multiplicación y la regla de los complementos pueden utilizarse juntas para simplificar en gran medida la solución a este tipo de problema: calcular la probabilidad de que entre varios ensayos, uno al menos dé algún resultado que se especificó. En casos como éste, es esencial que el significado del lenguaje se comprenda con claridad:

- "Uno al menos" equivale a "uno o más".
- El complemento de obtener uno al menos, de los elementos de un tipo en particular, es que usted no reciba elementos de ese tipo.

Suponga que una pareja planea tener tres hijos y quiere conocer la probabilidad de al menos una sea niña. Véanse las interpretaciones siguientes:

Una niña al menos entre tres hijos $=$ 1 o más niñas.

El complemento de "una niña al menos" $=$ no niñas $=$ los tres hijos son niños.

Calcularíamos esta probabilidad con facilidad realizando una lista del espacio muestral completo de ocho resultados, pero queremos ilustrar el uso de los complementos, ya que son útiles en muchos otros problemas que no se resuelven tan fácilmente.

EJEMPLO Género de hijos Calcule la probabilidad de que una pareja tenga al menos una niña entre tres hijos. Suponga que los niños y las niñas son igualmente probables, así como que el género de un hijo es independiente del género de cualquier hermano o hermana.

SOLUCIÓN

Paso 1: Use un símbolo para representar el suceso deseado. En este caso, sea A = al menos uno de los tres hijos es una niña.

Paso 2: Identifique el suceso que es el complemento de A.

$$\begin{aligned}\overline{A} &= \textit{no}\text{ tener al menos una niña entre tres hijos}\\ &= \text{todos los tres hijos son niños}\\ &= \text{niño y niño y niño}\end{aligned}$$

Paso 3: Calcule la probabilidad del complemento.

$$\begin{aligned}P(\overline{A}) &= P(\text{niño y niño y niño})\\ &= \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} = \frac{1}{8}\end{aligned}$$

Paso 4: Calcule $P(A)$ evaluando $1 - P(\bar{A})$.

$$P(A) = 1 - P(\bar{A}) = 1 - \frac{1}{8} = \frac{7}{8}$$

INTERPRETACIÓN Existen $7/8$ de probabilidad de que si una pareja tiene tres hijos, al menos uno sea una niña.

El principio que se utiliza en este ejemplo se resume como sigue:

Para calcular la probabilidad de *uno al menos* de algo, calcule la probabilidad de *ninguno*, y reste el resultado de 1. Esto es,

$$\boldsymbol{P(\text{uno al menos}) = 1 - P(\text{ninguno}).}$$

Probabilidad condicional

Ahora consideraremos el siguiente punto principal de esta sección, que se basa en el principio de que la probabilidad de un suceso suele verse afectada por el conocimiento previo de las circunstancias. Por ejemplo, si usted selecciona aleatoriamente a una persona de la población general, la probabilidad de obtener un hombre es de 0.5, pero si ya sabe que la persona a seleccionar cambia con frecuencia los canales de la televisión, con un control remoto, la probabilidad es de 0.999 (bueno, tal vez sea una pequeña exageración). Una *probabilidad condicional* de un suceso ocurre cuando la probabilidad se afecta por el conocimiento de otras circunstancias. La probabilidad condicional de que el suceso B ocurra, puesto que el suceso A ya ocurrió, se calcula usando la regla de la multiplicación [$P(A \text{ y } B) = P(A) \cdot P(B|A)$] y resolviendo para $P(B|A)$, así como dividiendo ambos lados de la ecuación entre $P(A)$.

Definición

Una **probabilidad condicional** de un suceso es una probabilidad que se obtiene con la información adicional de algún otro suceso que ya ocurrió. $P(B|A)$ denota la probabilidad condicional de que el suceso B ocurra, puesto que el suceso A ya ocurrió, y se calcula dividiendo la probabilidad de que ambos sucesos A y B ocurran entre la probabilidad del suceso A:

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

Esta fórmula es una expresión formal de la probabilidad condicional, pero recomendamos el siguiente método intuitivo.

Método intuitivo para la probabilidad condicional

La probabilidad condicional de B dado A, se calcula suponiendo que el suceso A ya ocurrió; bajo ese supuesto, se calcula la probabilidad de que el suceso B ocurra.

¿Coincidencias?

John Adams y Thomas Jefferson (el segundo y tercer presidentes de Estados Unidos) murieron el mismo día, el 4 de julio de 1826. El presidente Lincoln murió asesinado en el teatro Ford; el presidente Kennedy fue asesinado en un automóvil Lincoln hecho por la Ford Motor Company. Los sucesores a la presidencia, tanto de Lincoln como de Kennedy, fueron vicepresidentes con apellido Johnson. Catorce años *antes* del naufragio del *Titanic*, una novela describió el hundimiento del *Titán*, un barco que chocó con un iceberg; véase *The Wreck of the Titanic Foretold?*, de Martin Gardner. Gardner señala: "En casi todos los casos de coincidencias desconcertantes, es imposible hacer aunque sea una estimación burda de su probabilidad".

LA ESTADÍSTICA EN LAS NOTICIAS

¿Aplicación con resultado anticipado?

¿Se afecta la probabilidad de que lo acepten en una universidad si el aspirante opta por un resultado anticipado? Esta pregunta se trata con métodos de estadística; los resultados son algo sorprendentes. Al escribir acerca de las investigaciones del proceso de admisión a las universidades, la reportera Karen Arenson, del *New York Times*, afirma que "esto no sólo documenta que los estudiantes que hacen el examen pidiendo resultado anticipado tienen una clara ventaja sobre aquellos que no lo hacen —el equivalente de añadir 100 puntos en la calificación del SAT de un aspirante instantáneamente—, sino que también sugiere que el proceso es injusto, ya que muchos estudiantes de preparatorias menos prestigiadas no entienden la manera en que el sistema inclina las posibilidades de aceptación". Ella citó evidencia de 10 universidades con aspirantes que tienen calificaciones del SAT entre 1400 y 1490: al 70% de los estudiantes que solicitaron resultados anticipados se les aceptó, en comparación con el 51% de aceptación de quienes no solicitaron resultados anticipados.

PROBLEMA 3 DEL CAPÍTULO

EJEMPLO Ensayos clínicos de prueba de embarazo Remítase a la tabla 3-1, que se reproduce aquí para su conveniencia.

a. Si se elige aleatoriamente uno de los 99 sujetos, encuentre la probabilidad de que esa mujer pruebe positivo, ya que está embarazada.

b. Si se elige aleatoriamente uno de los 99 sujetos, encuentre la probabilidad de que ella esté embarazada, ya que la prueba resultó positiva.

Tabla 3-1 Resultados de prueba de embarazo

	Resultado de prueba positivo (indicó embarazo)	Resultado de prueba negativo (no indicó embarazo)
La mujer está embarazada	80	5
La mujer no está embarazada	3	11

SOLUCIÓN **a.** Queremos $P(\text{positivo} \mid \text{embarazada})$, la probabilidad de elegir a alguna mujer en quien la prueba fue positiva, *puesto que la mujer que se seleccionó estaba embarazada*. Aquí está el punto relevante: si suponemos que la mujer que se seleccionó estaba embarazada, estamos tratando sólo con las 85 mujeres del primer renglón de la tabla 3-1. De entre estas 85 mujeres, 80 dieron positivo; entonces,

$$P(\text{positivo} \mid \text{embarazada}) = \frac{80}{85} = 0.941$$

Puede calcularse el mismo resultado con la fórmula que se dio con la definición de probabilidad condicional. En los siguientes cálculos, utilizamos el hecho de que 80 de las 99 mujeres estaban embarazadas y dieron positivo. Además, 85 de las 99 mujeres estaban embarazadas. Tenemos

$$\begin{aligned} P(\text{positivo} \mid \text{embarazada}) &= \frac{P(\text{embarazada y positivo})}{P(\text{embarazada})} \\ &= \frac{80/99}{85/99} = 0.941 \end{aligned}$$

b. Aquí buscamos $P(\text{embarazada} \mid \text{positivo})$. Si suponemos que la mujer que se seleccionó dio positivo, estamos tratando con las 83 mujeres de la primera columna de la tabla 3-1. De entre estas 83 mujeres, 80 estaban embarazadas; entonces,

$$P(\text{embarazada} \mid \text{positivo}) = \frac{80}{83} = 0.964$$

Otra vez, se calcula el mismo resultado aplicando la fórmula para la probabilidad condicional:

$$\begin{aligned} P(\text{embarazada} \mid \text{positivo}) &= \frac{P(\text{positivo y embarazada})}{P(\text{positivo})} \\ &= \frac{80/99}{83/99} = 0.964 \end{aligned}$$

Comparando los resultados de los incisos *a* y *b*, veremos que $P(\text{positivo} \mid \text{embarazada})$ no es lo mismo que $P(\text{embarazada} \mid \text{positivo})$.

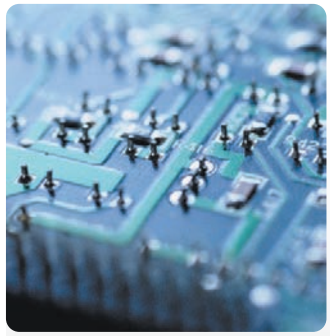
INTERPRETACIÓN El primer resultado, de $P(\text{positivo} \mid \text{embarazada}) = 0.941$, indica que una mujer embarazada tiene un 0.941 de probabilidad de dar positivo. Esto sugiere que si una mujer no da positivo, no puede confiar en que no está embarazada, así que tiene que continuar con pruebas adicionales. El segundo resultado, de $P(\text{embarazada} \mid \text{positivo}) = 0.964$, indica que para una mujer que dio positivo, hay un 0.964 de probabilidad de que ella esté realmente embarazada. Una mujer que dio positivo sería inteligente si se sometiera a pruebas adicionales.

Note que, en el ejemplo anterior, $P(\text{positivo} \mid \text{embarazada}) \neq P(\text{embarazada} \mid \text{positivo})$. Aunque los dos valores de 0.941 y 0.964 son muy cercanos en este ejemplo, dichos resultados estarían muy apartados en otros casos. El hecho de creer incorrectamente que $P(B \mid A) = P(A \mid B)$, suele llamarse *confusión del inverso*. Hay estudios que muestran que algunas veces los médicos dan información muy enredada cuando padecen de confusión del inverso.

Prueba de independencia

En la sección 3-4 establecimos que los eventos A y B son independientes si la ocurrencia de uno no afecta la probabilidad de la ocurrencia del otro. En la regla de la multiplicación para sucesos dependientes, si $P(B \mid A) = P(B)$; entonces, la ocurrencia del suceso A no tiene efecto en la probabilidad del suceso B y los dos sucesos A y B son independientes. Esto nos sugiere una prueba de independencia: si $P(B \mid A) = P(B)$, entonces A y B son sucesos independientes; sin embargo, si $P(B \mid A) \neq P(B)$, entonces A y B son sucesos dependientes. Otra prueba de independencia implica revisar si $P(A \text{ y } B)$ y $P(A) \cdot P(B)$ son iguales. Si las expresiones son iguales, los sucesos A y B son independientes. Si $P(A \text{ y } B) \neq P(A) \cdot P(B)$; entonces, A y B son sucesos dependientes. Estos resultados se resumen como sigue:

Dos sucesos A y B son *independientes* si	Dos sucesos A y B son *dependientes* si
$P(B \mid A) = P(B)$ o $P(A \text{ y } B) = P(A) \cdot P(B)$	$P(B \mid A) \neq P(B)$ o $P(A \text{ y } B) \neq P(A) \cdot P(B)$



Teorema de Bayes

Thomas Bayes (1702-1761) dijo que las probabilidades deben revisarse cuando averiguamos más sobre un suceso. He aquí una forma del teorema de Bayes:

$$P(A \mid B) = \frac{P(A) \cdot P(B \mid A)}{P(A) \cdot P(B \mid A) + P(\bar{A}) \cdot P(B \mid \bar{A})}$$

Suponga que el 60% de los circuitos integrados para computadora de una compañía se producen en una de sus fábricas (denotada por A) y el 40% en su otra fábrica (denotada por $\bar{A}$). Para un circuito integrado que se selecciona al azar, la probabilidad de que provenga de la fábrica A es de 0.60. Suponga además que se entera de que un circuito integrado está defectuoso y que las tasas de defectos para las dos fábricas son del 35% (para A) y del 25% (para $\bar{A}$). Se utiliza la fórmula anterior para determinar que hay una probabilidad de 0.677 de que el circuito integrado defectuoso provenga de la fábrica A.

3-5 Destrezas y conceptos básicos

Descripción de complementos. *En los ejercicios 1 a 4 haga una descripción escrita del complemento del suceso dado.*

1. **Prueba sanguínea** Cuando se prueba a 10 estudiantes para determinar su grupo sanguíneo, uno al menos tiene sangre del grupo A.
2. **Control de calidad** Cuando se envían 50 unidades de HDTV, todas están libres de defectos.

3. Auditorías del fisco Cuando un oficial del Tesoro selecciona 12 devoluciones de impuestos de ingresos y les hace auditoría, encuentra que ninguna de las devoluciones es correcta.

4. Éxito con las mujeres Cuando Mike pidió una cita a cinco mujeres diferentes, al menos una aceptó.

5. Probabilidad subjetiva Utilice la probabilidad subjetiva para estimar la probabilidad de seleccionar un adulto al azar y obtener una mujer, puesto que la persona que se seleccionó tiene el pelo más largo por 10 pulgadas. ¿Es la probabilidad suficientemente alta como para presumir que alguien con pelo largo casi con seguridad es mujer?

6. Probabilidad subjetiva Utilice la probabilidad subjetiva para estimar la probabilidad de seleccionar un adulto al azar y obtener un hombre, puesto que la persona que se seleccionó es dueña de una motocicleta. Si un investigador de crímenes encuentra que una motocicleta se registró a nombre de Pat Ryan, ¿será razonable creer que Pat es hombre?

7. Probabilidad de al menos una niña Si una pareja planea tener cinco hijos, ¿cuál es la probabilidad de que tenga al menos una niña? ¿Es dicha probabilidad lo suficientemente alta como para que la pareja sienta mucha confianza de que nacerá al menos una niña en cinco hijos?

8. Probabilidad de al menos una niña Si una pareja planea tener 12 hijos (llega a suceder), ¿cuál es la probabilidad de que nazca al menos una niña? Si la pareja eventualmente tuvo 12 hijos y todos fueron niños, ¿qué concluiría la pareja?

9. Al menos una multa de tránsito Si se pasa en un cruce que se equipó con una cámara de vigilancia, con la luz del semáforo en rojo, hay un 0.1 de probabilidad de recibir una multa de tránsito. Si usted se pasa este cruce cinco veces diferentes con la luz del semáforo en rojo, ¿cuál es la probabilidad de recibir al menos una multa de tránsito?

10. Al menos una respuesta correcta Si usted adivina al azar las respuestas a cuatro preguntas de opción múltiple (cada una con cinco respuestas posibles), ¿cuál es la probabilidad de tener al menos una correcta? Si un maestro poco exigente dice que para aprobar el examen es necesario al menos tener una respuesta correcta, ¿puede usted esperar razonablemente aprobar adivinando?

11. Probabilidad de una niña Calcule la probabilidad de que una pareja tenga una niña cuando nace su tercer hijo, puesto que los primeros dos hijos fueron niños. ¿Es el resultado igual a la probabilidad de que nazcan tres niñas entre tres hijos?

12. Genética de Mendel Remítase a la figura 3-3 en la sección 3-3, que incluye los chícharos que se usaron en un experimento genético. Si se selecciona aleatoriamente uno de los chícharos y se encuentra que tiene vaina verde, ¿cuál es la probabilidad de que tenga flor morada?

13. Pruebas clínicas de embarazo Remítase a la tabla 3-1 y suponga que una de las mujeres se selecciona al azar. Calcule la probabilidad de un resultado de prueba negativo, puesto que la mujer que se seleccionó no está embarazada. ¿Qué debe hacer una mujer si le aplican esta prueba de embarazo y obtiene un resultado negativo?

14. Pruebas clínicas de embarazo Remítase a la tabla 3-1 y suponga que una de las mujeres se selecciona al azar. Calcule la probabilidad de que la mujer que se seleccionó no esté embarazada, ya que la prueba indicó negativo. ¿El resultado es igual a la probabilidad de un resultado de prueba negativo, ya que la mujer que se seleccionó no está embarazada?

15. Redundancia en relojes despertadores Un estudiante pierde muchas clases por el mal funcionamiento de los relojes despertadores. En lugar de usar un reloj despertador, decide usar tres. ¿Cuál es la probabilidad de que al menos uno de sus relojes despertadores funcione correctamente si cada uno, por separado, tiene un 99% de probabilidad de funcionar correctamente? ¿En realidad gana mucho el estudiante usando tres relojes despertadores en lugar de uno solo?

16. Muestreo de aceptación Con un método del procedimiento que llaman *muestreo de aceptación*, se selecciona aleatoriamente y sin reemplazo una muestra de artículos. Tome en cuenta que el lote completo se rechazará si se encuentra al menos un defecto. La Niko Electronics Company acaba de fabricar 5,000 CD, de los cuales el 3% están defectuosos. Si se seleccionan 10 de los CD y se prueban, ¿cuál es la probabilidad de que se rechace el lote completo?

17. Uso de muestras de sangre compuestas Cuando se hacen pruebas de sangre para detectar infecciones por VIH, el procedimiento puede hacerse de forma más eficiente y menos costosa mezclando muestras de especímenes de sangre. Así, si las muestras de tres personas se combinan y la mezcla da un resultado negativo, sabemos que las tres muestras individuales son negativas. Calcule la probabilidad de un resultado positivo para tres muestras combinadas en una mezcla, suponiendo que la probabilidad de que una muestra de sangre individual dé positivo es de 0.1 (la probabilidad de la población "en riesgo" de acuerdo con datos del Departamento de Salud del estado de Nueva York).

18. Uso de muestras de agua compuestas El Departamento de Salud Pública del condado de Orange realiza pruebas al agua para determinar contaminación por la presencia de la bacteria *E. coli*. Con la finalidad de reducir costos de laboratorio, se mezclan las muestras de agua de seis áreas de natación públicas para efectuar una sola prueba, la cual sólo se hará más amplia si la muestra que se mezcla falla. Con base en resultados previos, hay un 2% de probabilidad de encontrar la bacteria *E. coli* en un área de natación pública. Calcule la probabilidad de que una muestra que se combinara de seis áreas de natación públicas revele la presencia de la bacteria *E. coli*.

Probabilidades condicionales. *En los ejercicios 19 a 22 use los datos de mortalidad que hubo en el* Titanic *de la tabla adjunta.*

	Hombres	Mujeres	Niños	Niñas
Sobrevivientes	332	318	29	27
Muertos	1360	104	35	18

19. Si seleccionamos aleatoriamente a una persona que abordó el *Titanic*, ¿cuál es la probabilidad de elegir un hombre, puesto que la persona que se seleccionó murió?

20. Si seleccionamos aleatoriamente a una persona que murió, ¿cuál es la probabilidad de elegir a un hombre?

21. ¿Cuál es la probabilidad de elegir a un niño o una niña, puesto que la persona que se seleccionó al azar es alguien que sobrevivió?

22. ¿Cuál es la probabilidad de elegir a un hombre o una mujer, ya que la persona seleccionada aleatoriamente es alguien que murió?

3-5 Más allá de lo básico

23. Montaña rusa La montaña rusa Rock'n'Roller de los estudios Disney-MGM, en Orlando, tiene dos asientos en cada una de sus 12 filas. Los pasajeros se asignan a los asientos en el orden en que van llegando. Si se sube a esta montaña rusa una vez, ¿cuál es la probabilidad de obtener el tan codiciado lugar de hasta adelante? ¿Cuántas veces habrá que subirse para tener un mínimo del 95% de probabilidad de que le toque el asiento delantero al menos una vez?

24. ¿Quién fue? La planta en Atlanta de la Medassist Pharmaceutical Company fabricó 400 marcapasos, de los cuales tres están defectuosos. La planta en Baltimore de la misma compañía fabricó 800 marcapasos, dos de los cuales salieron defectuosos. Si se selecciona al azar uno de los 1,200 marcapasos y se encuentra que está defectuoso, ¿cuál es la probabilidad de que lo hayan fabricado en Atlanta?

25. Uso de una tabla de dos entradas El Departamento de Salud del estado de Nueva York reporta una tasa de incidencia de VIH del 0.3% para la población general; en ciertas condiciones, las pruebas de investigación preliminar para el VIH son correctas un 95% de las veces (para verdaderos positivos y verdaderos negativos). Suponga que la población general conste de 100,000 personas.

a. Construya una tabla con un formato similar al de la tabla 3-1.

b. Usando la tabla del inciso *a* calcule $P(\text{VIH} \mid \text{positivo})$ para una persona seleccionada aleatoriamente de la población general. Esto es, calcule la probabilidad de seleccionar al azar a una persona con VIH, ya que esa persona dio positivo.

26. Fecha de cumpleaños compartida Calcule la probabilidad de que, de 25 personas que se seleccionaron al azar,

a. No haya dos que compartan la misma fecha de cumpleaños.

b. Al menos dos compartan la misma fecha de cumpleaños.

27. Monedas ocultas Un profesor de estadística lanza dos monedas que ningún estudiante logra ver. Un estudiante pregunta si una de las monedas cayó en cara. Puesto que la respuesta del profesor es sí, calcule la probabilidad de que ambas monedas cayeran en cara.

3-6 Probabilidades por medio de simulaciones

Los estudiantes que toman un curso introductorio de estadística suelen encontrar que el tema de la probabilidad es el más difícil. Algunos problemas de probabilidad pueden parecer simples, pero sus soluciones son increíblemente complejas. En este capítulo ya identificamos varias reglas básicas e importantes, las cuales suelen usarse para calcular probabilidades, pero en esta sección introducimos un enfoque muy diferente, que logra vencer gran parte de la dificultad que se encuentra en la aplicación de reglas formales. Este enfoque alternativo consiste en desarrollar una simulación.

Definición

Simulación de un procedimiento: proceso que se comporta de la misma forma que el procedimiento, de manera que se producen resultados similares.

Considere los ejemplos siguientes para entender mejor el uso de la simulación.

EJEMPLO **Selección del género** Cuando se prueban técnicas de selección del género, los investigadores médicos necesitan conocer valores de probabilidad de diferentes resultados, como, por ejemplo, la probabilidad de tener al menos 60 niñas entre 100 niños. Suponiendo que el nacimiento de un hombre o de una mujer es igualmente probable, describa una simulación que dé como resultado los géneros de 100 bebés recién nacidos.

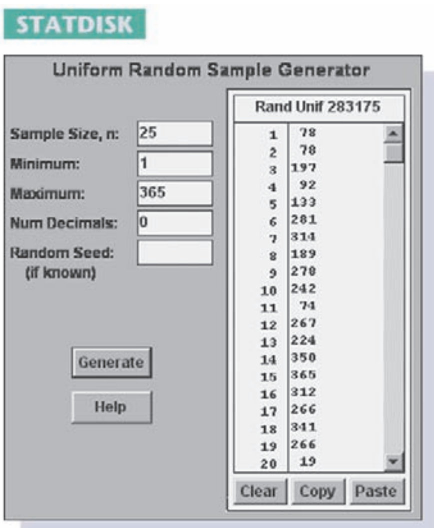
SOLUCIÓN Una opción es simplemente lanzar una moneda al aire 100 veces, con la cara representando a las mujeres, y la cruz, a los hombres. Otra opción es usar una calculadora o computadora para generar aleatoriamente ceros y unos (el 0 representa a un niño y el 1 representa a una niña). Los números deben generarse de forma que sean igualmente probables.

EJEMPLO Las mismas fechas de cumpleaños El ejercicio 26 en la sección 3-5 se refiere al clásico problema de fecha de cumpleaños, en el que encontramos la probabilidad de que, en un grupo 25 personas que se seleccionaron al azar, al menos dos compartan la misma fecha de cumpleaños. La solución teórica es difícil, ya que resulta poco práctico encuestar a muchos grupos diferentes de 25 personas, por lo que desarrollamos una simulación.

SOLUCIÓN Comience por representar fechas de cumpleaños con números enteros del 1 a 365, donde 1 = 1 de enero, 2 = 2 de enero, . . . , 365 = 31 de diciembre. Después, use una calculadora o un programa de cómputo para generar 25 números aleatorios, todos entre 1 y 365. Estos números pueden ordenarse, ya que así es fácil estudiar la lista para determinar si dos de las fechas de cumpleaños que se simularon son iguales. Es posible repetir el proceso tantas veces como queramos, hasta quedar satisfechos de tener bases firmes para determinar la probabilidad. Nuestro estimado de la probabilidad es el número de veces que tuvimos al menos dos fechas de cumpleaños iguales, dividido entre el número total de grupos de 25 que se generaron.

Hay varias maneras de obtener números de 1 a 365 generados aleatoriamente, incluyendo la siguiente:

- **Una tabla de números aleatorios:** Remítase, por ejemplo, al *CRC Standard Probability and Statistics Tables and Formulae*, que contiene una tabla de 14,000 dígitos. (Existen muchas formas de extraer números de 1 a 365 en este tipo de tablas. Una consiste en tomar los dígitos en las primeras columnas, ignorando 000 y cualquier número mayor a 365).
- **STATDISK:** Seleccione **Data** de la barra del menú principal, luego seleccione **Uniform Generator**; después, proceda a introducir un tamaño muestral de 25, un mínimo de 1 y un máximo de 365; introduzca 0 para el número de lugares decimales. La pantalla que resulta en el STATDISK se reproduce abajo. Use **copy/paste**, para copiar el conjunto de datos al **Sample Editor**, donde los valores pueden acomodarse en orden creciente. En la pantalla del STATDISK que se presenta aquí, vemos que las primeras dos personas tienen la misma fecha de cumpleaños, que es el día 78 del año.



Monos mecanógrafos

Una aseveración clásica dice que un mono que golpea al azar un teclado, tarde que temprano produciría las obras completas de Shakespeare, suponiendo que continúe tecleando siglo tras siglo. Se utilizó ya la regla de la multiplicación para probabilidades con la finalidad de obtener estimados de esta clase. Algunos consideran muy pequeño un resultado de 1,000,000,000,000,000,000,000,000,000,000,000,000 años. Con algo similar en mente, sir Arthur Eddington escribió este poema: "Había una vez un sesudo babuino, que soplaba y soplaba un fagot; 'pues estoy convencido —decía—, de que en miles de millones de años, si sigo, al soplar me saldrá una canción'".

LA ESTADÍSTICA EN LAS NOTICIAS

Para ganar, apueste con audacia

El diario *New York Times* publicó un artículo de Andrew Pollack en el cual se reportó que el casino Mirage en Las Vegas tenía menores ganancias que las que se esperaban. Él escribió que "las ganancias del Mirage pueden ser particularmente volátiles, ya que se favorece a los grandes apostadores, jugadores que llegan a apostar $100,000 o más en una mano de cartas. La ley de los promedios no funciona con tanta consistencia para unas cuantas apuestas grandes como lo hace para miles de pequeñas. . .". Esto refleja el principio más fundamental al apostar: para ganar, ¡ponga una apuesta grande en lugar de muchas apuestas chicas! Con el juego adecuado, por ejemplo el de dados, usted tiene poco menos del 50% de posibilidades de duplicar su dinero si se anima a una apuesta grande. Al hacer muchas apuestas pequeñas, la probabilidad de duplicar su dinero disminuye sustancialmente.

- **Minitab:** Seleccione **Calc** en la barra del menú principal; después, seleccione **Random Data** e **Integer**. En el cuadro de diálogo, introduzca 25 para el número de filas, guarde los resultados en la columna C1; tras esto, ingrese un mínimo de 1 y un máximo de 365. Entonces ya está listo para usar **Manip** y **Sort** para acomodar los datos en orden creciente. El resultado se verá como se muestra abajo, pero los números no serán los mismos. Este resultado del Minitab de 25 números indica que el 9o y el 10o son iguales.

Minitab

	C1	C2
↓		
1	38	
2	48	
3	59	
4	71	
5	101	
6	107	
7	122	
8	129	
9	153	
10	153	
11	163	

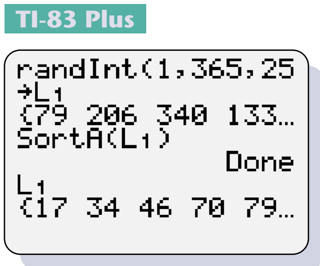
- **Excel:** Haga clic en la celda que se encuentra en la esquina superior izquierda, después haga clicen el icono de función **fx.** Seleccione **Matemáticas y Trigonometría**, y después seleccione **RANDBETWEEN**. En el cuadro de diálogo, escriba 1 para el límite inferior (*bottom*) y ponga 365 como límite

Excel

	A	
1	15	
2	3	
3	15	
4	362	
5	164	
6	184	
7	158	
8	59	
9	143	
10	85	
11	134	

superior (*top*). Después de obtener el número aleatorio en la primera celda, haga clic y mantenga presionado el botón para arrastrar la esquina inferior derecha de la primera celda, luego arrástrela hacia abajo hasta resaltar 25 celdas. Cuando suelte el botón, deben aparecer los 25 números aleatorios. La pantalla que se reproduce aquí, indica que los números 1o y 3o son iguales.

- **Calculadora TI-83 Plus:** Oprima la tecla **MATH**, seleccione **PRB**, luego escoja **randInt** y proceda a introducir el mínimo de 1, el máximo de 365 y 25 para el número de valores. Esto es, introduzca randInt(1,365,25). Observe la pantalla de la TI-83 Plus, la cual indica el uso de **randInt** para generar los números; luego se guardaron en la lista L1, donde se sortearon y se mostraron. La imagen de la pantalla que se reproduce aquí indica que no hay números iguales entre los pocos que se alcanzan a ver. Para observar la lista completa de números generados, oprima **STAT** (Estadística) y seleccione **Edit.**



Es en extremo importante construir una simulación que se comporte precisamente como el procedimiento real. En el siguiente ejemplo demostramos una forma correcta y una forma incorrecta de construir una simulación.

EJEMPLO Simulación de dados Describa un procedimiento para simular el acto de tirar un par de dados.

SOLUCIÓN En el procedimiento de tirar un par de dados, cada uno de los dos dados nos dará un número entre 1 y 6 (inclusive); estos dos números se suman. Cualquier simulación debe efectuar lo mismo. Hay una manera correcta y una incorrecta de simular un tiro de dos dados.

La manera correcta: Generar aleatoriamente un número entre 1 y 6, generar aleatoriamente otro número entre 1 y 6; luego, sumar los dos resultados.

La manera incorrecta: Generar aleatoriamente números entre 2 y 12. Este procedimiento es similar a tirar dos dados, en el sentido de que los resultados dan siempre entre 2 y 12, pero estos resultados entre 2 y 12 son igualmente probables. Con dados reales, los valores entre 2 y 12 *no son* igualmente probables. Esta simulación produciría muchos resultados confusos.

Algunos problemas de probabilidad se resuelven sólo por estimación de la probabilidad utilizando observaciones reales o construyendo una simulación. La extensa disponibilidad de calculadoras y computadoras facilita mucho el uso de métodos de simulación, tanto que ahora las simulaciones se emplean con frecuencia para determinar valores de probabilidad.

La secretaria aleatoria

Un clásico problema de probabilidad dice así: una secretaria prepara 50 cartas distintas y las dirige a 50 personas diferentes, pero las revuelve aleatoriamente antes de meterlas en los sobres. ¿Qué probabilidad hay de que al menos una carta quede en el sobre que le corresponde? Aunque podría parecer que la probabilidad es pequeña, en realidad es de 0.632. Incluso con un millón de cartas y un millón de sobres, la probabilidad es de 0.632. La solución está mucho más allá del alcance de este texto.

3-6 Destrezas y conceptos básicos

En los ejercicios 1 a 8 utilice la lista de números que se generaron aleatoriamente y se encuentra al margen. (Se obtiene una lista similar usando calculadoras, computadoras, resultados de la lotería o tablas especiales de números aleatorios).

1. Simulación de respuestas por adivinación Suponga que usted quiere usar los dígitos de la lista adjunta para simular adivinaciones en un examen de verdadero/falso. Si un dígito impar representa “verdadero” y un dígito par representa “falso”, haga una lista de cinco respuestas correspondientes con la primera fila de dígitos.

2. Simulación de dados Suponga que usted quiere usar los dígitos de la lista adjunta para simular el hecho de tirar un solo dado. Si se usan los dígitos 1, 2, 3, 4, 5, 6, mientras se ignoran todos los demás, haga una lista de los resultados que se obtuvieron con las primeras dos filas.

46196
99438
72113
44044
86763
00151
64703
78907
19155
67640
98746
29910
82855
25259
14752
85446
75260
92532
87333
55848

3. Simulación de fabricación La compañía Telektronic está experimentando con un nuevo proceso de fabricación de fusibles, en la cual la tasa de defectos es del 20%. Es posible simular fusibles defectuosos usando 0 y 1, mientras que 2, 3, 4, 5, 6, 7, 8, 9 representan fusibles en buen estado (de manera que se consideran el 20% de defectos). Identifique los fusibles aceptables y defectuosos que corresponden con la primera fila de dígitos.

4. Simulación de fechas de cumpleaños En un ejemplo de esta sección se señaló que las fechas de cumpleaños pueden simularse generando enteros entre 1 y 365. Si usamos entradas en una lista de dígitos aleatorios, se representa el 1 de enero como 001, el 2 de enero como 002, . . . , y el 31 de diciembre como 365. Todos las demás ternas de dígitos deben ignorarse. Utilizando este método, la primera fila nos da la fecha de cumpleaños válida de 196. Haga una lista de las siguientes cinco fechas de cumpleaños que se logran obtener de esta forma.

5. Simulación de familias con cinco hijos Use los dígitos aleatorios al margen para desarrollar una simulación con la finalidad de calcular la probabilidad de tener al menos dos niñas en una familia de cinco hijos. Describa la simulación y luego estime la probabilidad con base en sus resultados. ¿De qué forma es posible comparar el resultado con la cifra correcta de 0.813? (*Sugerencia:* Haga que los dígitos impares representen niñas).

6. Simulación de tres dados Use los dígitos aleatorios al margen y desarrolle una simulación para el tiro de tres dados. Describa la simulación, luego proceda a estimar la probabilidad de obtener un total de 10 cuando se tiran tres dados. ¿El resultado es comparable con el resultado correcto de 0.125? (*Sugerencia:* Use sólo los dígitos 1, 2, 3, 4, 5, 6, e ignore todos los demás dígitos).

7. Simulación de zurdos El 10% de las personas son zurdas. En un estudio de destreza, se seleccionan al azar grupos de cinco. Utilice los dígitos aleatorios al margen y desarrolle una simulación para calcular la probabilidad de obtener al menos una persona zurda en un grupo de cinco. ¿El resultado es comparable con el resultado correcto de 0.410, que se puede calcular usando las reglas de probabilidad de este capítulo? (*Sugerencia:* Puesto que el 10% de las personas son zurdas, deje que el dígito 0 represente a alguien que es zurdo y que los otros dígitos representen a alguien que no es zurdo).

8. Simulación de hibridación Cuando Mendel realizó sus famosos experimentos de hibridación, utilizó chícharos con vainas verdes y vainas amarillas. Un experimento incluyó la mezcla de chícharos de tal forma que se esperaba que el 25% de los chícharos vástagos tuvieran vainas amarillas. Use los dígitos aleatorios al margen y desarrolle una simulación para calcular la probabilidad de que cuando se produzcan dos chícharos vástagos, al menos uno de ellos contenga vainas amarillas. ¿El resultado es comparable con la probabilidad correcta de 7/16, que puede calcularse usando las reglas de probabilidad de este capítulo? (*Sugerencia:* Puesto que se espera que el 25% de los vástagos tengan vainas amarillas y el 75% tengan vainas verdes, haga que el dígito 1 represente vainas amarillas y que los dígitos 2, 3, 4 representen vainas verdes; ignore cualquier otro dígito).

T *En los ejercicios 9 a 12 desarrolle una simulación utilizando la calculadora TI-83 Plus, el STATDISK, el Minitab, Excel, o cualquier otro programa o calculadora adecuados.*

9. Simulación de familias de cinco hijos En el ejercicio 5 utilizamos los dígitos al margen para estimar la probabilidad de tener al menos dos niñas en una familia de cinco hijos. En lugar de usar los mismos dígitos, desarrolle su propia simulación para calcular la probabilidad de que haya al menos dos niñas en una familia de cinco hijos. Simule 100 familias. Describa la simulación y después estime la probabilidad con base en los resultados.

10. Simulación de tres dados En el ejercicio 6 utilizamos los dígitos al margen para simular el tiro de dados. En lugar de usar los mismos dígitos, desarrolle su propia simulación para tirar tres dados. Simule tirar tres dados 100 veces. Describa la simulación y maneje el resultado para estimar la probabilidad de tener un total de 10 cuando se tiran tres dados.

11. Simulación de zurdos En el ejercicio 7 utilizamos los dígitos al margen para simular personas que son zurdas o diestras. (El 10% de las personas son zurdas). Desarrolle una simulación para calcular la probabilidad de obtener al menos una persona zurda en un grupo de cinco. Simule 100 grupos de cinco.

12. Simulación de hibridación En el ejercicio 8 utilizamos los dígitos al margen como base para simular la hibridación de chícharos. Suponga otra vez que se espera que el 25% de los chícharos vástagos tengan vainas amarillas, pero desarrolle su simulación y genere 100 pares de vástagos. Con base en sus resultados, estime la probabilidad de tener al menos un chícharo con vainas verdes cuando se obtienen dos chícharos vástagos.

3-6 Más allá de lo básico

13. Simulación del problema de Monty Hall Un problema que ha atraído gran atención es el problema de Monty Hall, que se inspiró en el antiguo programa de concurso de televisión “Let’s Make a Deal”, que presenta Monty Hall. Suponga que usted es un concursante que eligió una de tres puertas, después de que se le dijo que detrás de dos de ellas no hay nada, pero que detrás de una de las tres está un Corvette rojo último modelo. Entonces, se le da la opción de quedarse con su primera selección o cambiarla. ¿Debe quedarse con su primera elección o le conviene cambiar? (De acuerdo con la revista *Chance*, las escuelas de negocios de instituciones como Harvard y Stanford usan este problema para ayudar a los estudiantes a relacionarse con la toma de decisiones).

14. Simulación de fechas de cumpleaños
   a. Elabore una simulación para calcular la probabilidad de que, cuando 50 personas se seleccionan al azar, al menos dos de ellas tengan la misma fecha de cumpleaños. Describa la simulación y estime la probabilidad.
   a. Elabore una simulación para calcular la probabilidad de que, cuando 50 personas se seleccionan al azar, al menos tres de ellas tengan la misma fecha de cumpleaños. Describa la simulación y estime la probabilidad.

15. Genética: simulación de control poblacional Un clásico problema de probabilidad se refiere a un rey que quería incrementar la proporción de mujeres decretando que después de que una madre diera a luz a un hijo hombre, se le prohibiera tener más hijos. El rey razona que algunas familias tendrán sólo un varón, mientras que en otras habrá unas cuantas mujeres y un hombre, luego de lo cual la proporción de niñas se incrementará. ¿Es correcto su razonamiento? ¿Se incrementará la proporción de niñas?

3-7 Conteo

¿Cuál es la probabilidad de que usted gane en la lotería? En la lotería de Maine, una lotería típica, usted debe escoger seis números entre 1 y 42, inclusive. Si elige la misma combinación de seis números que los oficiales de la lotería sacan al azar, ganará el premio mayor, que a veces es de millones de dólares. Hay algunos premios menores, pero son relativamente insignificantes. Utilizando el enfoque clásico para la probabilidad (puesto que los resultados son igual de probables), la probabilidad de ganar la lotería se encuentra usando $P(\text{ganar}) = s/n$, donde s es el número de formas en que usted puede ganar y n es el número total de resultados posibles. Con la lotería de Maine $s = 1$, puesto que sólo existe una manera de ganar el premio mayor: escoger la misma combinación de seis números que se saca en la lotería. Sabiendo que sólo hay una manera de ganar, ahora necesitamos calcular n, el número total de resultados, es decir, ¿cuántas combinaciones de seis números son posibles cuando selecciona números de 1 a 42? Escribir una lista de las posibilidades tomaría alrededor de un año de trabajo sin parar; además, ese método no le dejaría tiempo para estudiar estadística. Necesitamos una manera más práctica de calcular el número total de posibilidades. Esta sección introduce métodos eficientes para calcular números de ese tipo, sin hacer listas directamente y contar las posibilidades. Regresaremos a dicho problema de la lotería después de presentar algunos principios básicos. Comencemos por la *regla fundamental de conteo*.

Regla fundamental de conteo

Para una secuencia de dos sucesos en la que el primer suceso puede ocurrir de m formas y el segundo suceso puede ocurrir de n formas, los sucesos juntos pueden ocurrir un total de $m \cdot n$ formas.

La regla fundamental de conteo se extiende fácilmente a situaciones que impliquen más de dos eventos, como se ilustra en los ejemplos siguientes.

EJEMPLO Las bases del robo Los sistemas comunes de alarma para casas tienen un código que consta de cuatro dígitos. Los dígitos (0 hasta 9) pueden estar repetidos, aunque deben ingresarse en el orden correcto. Suponga que usted planea tener acceso intentando códigos hasta encontrar el correcto. ¿Cuántos códigos diferentes son posibles?

SOLUCIÓN Hay 10 valores posibles para cada uno de los cuatro dígitos; entonces, el número de códigos posibles distintos es de $10 \cdot 10 \cdot 10 \cdot 10 = 10{,}000$. Aunque los 10,000 códigos pueden intentarse en alrededor de 11 horas, los sistemas de alarma normalmente se diseñaron para que el sistema rechace intentos subsecuentes después de unas cuantas entradas incorrectas. Además de los problemas morales y legales de ser ladrón profesional, parece que hay un asunto matemático que también sugiere hacer otra carrera.

EJEMPLO Cotinina en fumadores El conjunto de datos 6 del Apéndice B lista niveles de cotinina que se midieron en una muestra de personas de cada uno de los tres grupos: fumadores (denotados aquí por F), no fumadores que están expuestos al humo del tabaco (denotados por E), y no fumadores que no están expuestos al humo del tabaco (denotados por N). Cuando el cuerpo absorbe la nicotina, se produce cotinina. Si calculamos el promedio del nivel de cotinina de cada uno de los tres grupos y luego acomodamos dichos promedios en orden creciente, obtendremos la secuencia de sucesos NEF. Un cabildero en contra del tabaquismo afirma que esto es evidencia de que consumir tabaco daña la salud, porque la presencia de cotinina se incrementa a medida que la exposición y el uso del tabaco se incrementan. ¿De cuántas formas pueden acomodarse los tres grupos que se denotan con N, E y F? Si se selecciona al azar un arreglo, ¿cuál es la probabilidad de tener la secuencia NEF? ¿Es la probabilidad lo suficientemente baja como para concluir que la secuencia NEF indica que la presencia de cotinina se incrementa a medida que la exposición y el uso del tabaco también se incrementan?

SOLUCIÓN Al hacer arreglos de secuencias de los grupos N, E y F, hay tres posibles opciones para el primer grupo, dos opciones para el segundo grupo y sólo una opción para el tercer grupo. El número total de arreglos posibles es entonces

$$3 \cdot 2 \cdot 1 = 6$$

Existen seis maneras diferentes de acomodar los grupos N, E y F (que pueden listarse como NEF, NFE, EFN, ENF, FNE y FEN). Si seleccionamos aleatoriamente una de las seis secuencias posibles, la probabilidad de obtener la secuencia NEF es de $1/6$. Puesto que la probabilidad de $1/6$ es relativamente alta, sabemos que la secuencia NEF puede ocurrir con facilidad por posibilidad. La probabilidad no es suficientemente baja como para concluir que la secuencia NEF indique que la presencia de cotinina se incrementa a medida que la exposición y uso del tabaco también lo hacen. Necesitaríamos tener una probabilidad más baja; por ejemplo, de 0.01.

En el ejemplo anterior, encontramos que tres grupos pueden acomodarse en $3 \cdot 2 \cdot 1 = 6$ formas diferentes. Esta solución específica se generaliza utilizando la siguiente notación para el símbolo ! y la siguiente *regla factorial*.

Notación

El **símbolo factorial !** denota el producto de números enteros positivos decrecientes. Por ejemplo, $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$. Por definición especial, $0! = 1$. (Muchas calculadoras traen una tecla factorial. En la calculadora TI-83 Plus, ingrese primero el número, luego presione **MATH**; luego, seleccione **PRB** Y el elemento 4 del menú; por último, presione la tecla **ENTER**).

¿Cuántas veces hay que barajar?

Después de extensas investigaciones, el matemático de Harvard, Persi Diaconis encontró que se necesita barajar siete veces un mazo de naipes para obtener un mezclado completo. La mezcla es completa en el sentido de que todos los acomodos posibles de los naipes son igualmente probables. Barajar más de siete veces no surtirá un efecto significativo, en tanto que menos de siete no será suficiente. Los repartidores de naipes en los casinos rara vez barajan los mazos siete veces o más, así que los mazos no se mezclan totalmente. Algunos jugadores expertos aprovechan las mezclas incompletas que resultan de barajar menos de siete veces.

Elección de códigos de seguridad

Utilizamos códigos de seguridad personales para tener acceso a máquinas contestadoras telefónicas, cuentas de Internet de computadora y sistemas de seguridad para casas, entre otros sistemas. La seguridad de tales códigos depende del gran número de posibilidades diferentes, aunque ahora los piratas informáticos cuentan con sofisticadas herramientas capaces de superar este obstáculo con creces. Los investigadores encontraron ya que usando variaciones del nombre y apellido del usuario, además de otros 1,800 nombres, identificarían del 10% al 20% de las contraseñas de sistemas de cómputo típicos. Cuando escoja una contraseña, *no use* variaciones de ningún nombre, ni una palabra del diccionario, ni una palabra más corta que siete caracteres, ni números telefónicos, ni números del Seguro Social. Incluya caracteres no alfabéticos, como números o símbolos de puntuación.

Regla factorial

Una colección de n elementos distintos se puede acomodar de $n!$ diferentes maneras (esta **regla factorial** refleja el hecho de que el primer elemento se puede seleccionar de n maneras distintas, el segundo de $n - 1$ maneras y así sucesivamente).

Los problemas de ruta con frecuencia implican la aplicación de la regla factorial. AT&T quiere hacer sus llamadas telefónicas a través de las redes más cortas. Federal Express desea encontrar las rutas más cortas para sus entregas. American Airlines busca la ruta más corta para regresar a los miembros de la tripulación a sus casas. Véase el siguiente ejemplo.

EJEMPLO Rutas a 50 capitales Por su éxito en el curso de estadística, a usted lo contrató la organización Gallup. Su primer trabajo consiste en realizar una encuesta en cada una de las 50 capitales de los estados de Estados Unidos. Como usted se encuentra planeando su ruta de viaje, quiere determinar el número de rutas diferentes posibles. ¿Cuántas rutas diferentes son posibles?

SOLUCIÓN Aplicando la regla factorial sabemos que 50 elementos pueden acomodarse en 50! formas diferentes. Es decir, a las 50 capitales de estado es posible acomodarlas en 50! formas, o sea

$$30{,}414{,}093{,}201{,}713{,}378{,}043{,}612{,}608{,}166{,}064{,}768,$$
$$844{,}377{,}641{,}568{,}960{,}512{,}000{,}000{,}000{,}000$$

Ahora sí tenemos un número grande.

El ejemplo anterior es una variación del clásico problema que se conoce como *problema del vendedor viajero*, que es especialmente interesante, pues el gran número de posibilidades existentes significa que no estamos en condiciones de utilizar una computadora para calcular la distancia de cada ruta. El tiempo que tomaría calcular la ruta más corta, aun con la computadora más rápida, es de alrededor de

$$1{,}000{,}000{,}000{,}000{,}000{,}000{,}000{,}000{,}000{,}000{,}000{,}000{,}000{,}000 \text{ siglos.}$$

Constantemente se dedican esfuerzos considerables para tratar de encontrar métodos eficientes que resuelvan problemas de este tipo.

De acuerdo con la regla factorial, n diferentes elementos pueden ser acomodados de $n!$ diferentes maneras. Algunas veces tenemos n elementos diferentes, pero necesitamos seleccionar sólo algunos de ellos en lugar de todos. Si hay que realizar encuestas en capitales estatales, como en el ejemplo anterior, pero sólo tenemos tiempo de visitar cuatro capitales, el número de posibles rutas diferentes es de $50 \cdot 49 \cdot 48 \cdot 47 = 5{,}527{,}200$. Otra forma de obtener este mismo resultado es evaluar

$$\frac{50!}{46!} = 50 \cdot 49 \cdot 48 \cdot 47 = 5{,}527{,}200$$

En este cálculo, advierta que al dividir el número factorial del numerador entre el número factorial del denominador, sólo permanecen los factores de 50, 49, 48 y 47.

Se generaliza este resultado observando que si tenemos n elementos disponibles diferentes y queremos seleccionar un número r de ellos, el número de combinaciones es de $n!/(n-r)!$ como en $50!/46!$ Dicha generalización se conoce comúnmente como *regla de las permutaciones*.

Regla de las permutaciones (cuando todos los elementos son diferentes)

El número de **permutaciones** (o secuencias) de r elementos que se seleccionan entre n elementos disponibles (sin reemplazo) es

$$_nP_r = \frac{n!}{(n-r)!}$$

Muchas calculadoras logran evaluar expresiones de $_nP_r$.

Es muy importante reconocer que la regla de las permutaciones requiere las siguientes condiciones:

- Debemos tener un total de n *diferentes* elementos disponibles. (Esta regla no se aplica si algunos de los elementos son idénticos a otros).
- Debemos seleccionar r entre los n elementos (sin reemplazo).
- Debemos considerar que los reacomodamientos de los mismos elementos son secuencias diferentes.

Cuando utilizamos los términos *permutaciones, acomodos* o *secuencias,* implicamos que *se toma en cuenta el orden,* en el sentido de que diferentes ordenamientos de los mismos elementos cuentan como secuencias distintas. Las letras ABC se pueden acomodar de seis formas distintas: ABC, ACB, BAC, BCA, CAB y CBA. (Más adelante nos referiremos a las combinaciones, en las cuales tales acomodos no se consideran distintos). En el ejemplo siguiente se nos pide calcular el número total de secuencias distintas posibles. Eso sugiere el uso de la regla de las permutaciones.

EJEMPLO Programación de televisión Usted acaba de ser contratado para conformar la programación de la cadena de televisión Fox. Cuando está seleccionando los programas a transmitir el lunes por la noche, encuentra que tiene 27 programas disponibles y que debe seleccionar cuatro de ellos. El orden de los programas es importante, por los efectos de liderazgo. ¿Cuántas secuencias diferentes de cuatro programas son posibles cuando hay 27 programas disponibles?

SOLUCIÓN Necesitamos seleccionar $r = 4$ programas de $n = 27$ que están disponibles.

$$_nP_r = \frac{n!}{(n-r)!} = \frac{27!}{(27-4)!} = 421{,}200$$

Hay 421,200 arreglos posibles diferentes de cuatro programas que se seleccionaron de entre 27 disponibles.

Ganar centavos en la lotería

Muchas personas gastan grandes cantidades de dinero comprando billetes de lotería, a pesar de no tener un sentido realista de sus oportunidades de ganar. El hermano Donald Kelly, del Colegio Marista, propone esta analogía: ¡ganar la lotería es equivalente a recoger correctamente el centavo "ganador" de una columna de centavos que tiene una altura de 21 millas! Los aviones comerciales por lo regular vuelan a una altitud de seis millas, así que trate de imaginar una columna de centavos de una altura de más del triple de la que alcanzan esos jets de altos vuelos; además, imagínese escogiendo el centavo de esa columna que representa un billete de lotería ganador. Usando los métodos de esta sección, calcule la probabilidad de ganar la lotería de su estado y luego determine la altura de la columna de centavos correspondiente.

Escasez de números telefónicos

Las compañías telefónicas con frecuencia separan regiones con un código de área por regiones con dos o más códigos de área, porque las nuevas líneas de fax e Internet están próximas a agotar los números que pueden suscribirse bajo un solo código. Puesto que un número telefónico de siete dígitos no es posible que comience con 0 o 1, hay $8 \cdot 10 \cdot 10 \cdot 10 \cdot 10 \cdot 10 \cdot 10 = 8{,}000{,}000$ de números telefónicos diferentes posibles.

Antes de los teléfonos celulares, los equipos de fax y la Internet, todos los números gratuitos tenían un prefijo de 800. Estos números 800 duraron 29 años antes de que todos se asignaran. Se introdujo el prefijo 888 con el objetivo de contribuir a satisfacer la demanda de números gratuitos, pero se estimó que pasarían sólo 2.5 años para que los números 888 se agotaran. Lo que sigue a futuro: números gratuitos con el prefijo 877. Las técnicas de conteo de esta sección se usan para determinar la cantidad de números gratuitos distintos posibles con un prefijo dado, ya que es necesario satisfacer las demandas del futuro.

En ocasiones necesitamos calcular el número de permutaciones cuando algunos de los elementos son idénticos a otros. La siguiente variación de la regla de las permutaciones se aplica a tales casos.

Regla de las permutaciones (cuando algunos elementos son idénticos a otros)

Si hay n elementos con n_1 iguales, n_2 iguales, . . . , n_k iguales, el número de permutaciones de los n elementos es

$$\frac{n!}{n_1!n_2! \cdots n_k!}$$

EJEMPLO Invertir en acciones Los ejemplos clásicos de la regla de permutaciones son aquellos que muestran, por ejemplo, que las letras de la palabra Mississippi es posible acomodarlas de 34,650 formas diferentes o que las letras de la palabra estadística lo serían de 2,494,800 formas. En lugar de ello, consideraremos las letras *DDDDDEEEE*, que representan una secuencia de años recientes en los que el promedio industrial Dow Jones estuvo por debajo (*D*) de la media o por encima (*E*) de la media. ¿De cuántas maneras se acomodarían las letras *DDDDDEEEE*? ¿Parece que la secuencia es aleatoria? ¿Existe un patrón que sugiera que sería prudente invertir en acciones?

SOLUCIÓN En la secuencia *DDDDDEEEE* tenemos $n = 9$ elementos, con $n_1 = 5$ iguales y otros $n_2 = 4$ iguales. El número de permutaciones se calcula como sigue:

$$\frac{n!}{n_1!n_2!} = \frac{9!}{5!4!} = \frac{362{,}880}{2880} = 126$$

Hay 126 diferentes formas en que las letras *DDDDDEEEE* pueden acomodarse. Ya que hay 126 diferentes posibles arreglos y sólo dos de ellos (*DDDDDEEEE* y *EEEEDDDDD*) resultan en todas las letras agrupadas juntas, parece que la secuencia no es aleatoria. Puesto que todos los valores por *debajo* suceden al principio y todos los valores por *encima* al final, parece haber un patrón de incremento del valor de las acciones. Esto sugiere que sería prudente invertir en acciones. (Véase en la sección 13-7 la *prueba de rachas para detectar aleatoriedad*, un procedimiento formal que se usa con frecuencia para identificar tendencias económicas).

El ejemplo anterior comprende n elementos, cada uno perteneciente a una de dos categorías. Cuando sólo hay dos categorías, es posible estipular que x de los elementos son iguales y que los otros $n-x$ elementos también son iguales; entonces, la fórmula de las permutaciones se simplifica a

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

Este resultado, en particular, se usará para el análisis de probabilidades binomiales, que se explica en la sección 4-3.

Cuando tenemos la intención de seleccionar r elementos de entre n elementos diferentes, *sin tomar en cuenta el orden*, nos preocupan en realidad las combinaciones posibles, más que las permutaciones. Es decir, **cuando diferentes ordenamientos de los mismos elementos cuentan por separado, tenemos un problema de permutaciones, pero cuando los diferentes ordenamientos de los mismos elementos no cuentan por separado, nos enfrentamos a un problema de combinaciones;** en tal caso se aplica la regla siguiente:

Regla de las combinaciones

El número de **combinaciones** de r elementos que se seleccionaron de entre n elementos diferentes es

$$_nC_r = \frac{n!}{(n-r)!r!}$$

Muchas calculadoras se diseñaron para evaluar $_nC_r$.

Es muy importante reconocer que, en la aplicación de la regla de las combinaciones, se aplican las siguientes condiciones:

- Debemos tener un total de n elementos diferentes disponibles.
- Debemos seleccionar r de los n elementos (sin reemplazo).
- Debemos considerar que los reacomodos de tales elementos son los mismos. (La combinación *ABC* es la misma que *CBA*).

Puesto que quizá resulte confuso escoger entre la regla de las permutaciones y la regla de las combinaciones, damos el ejemplo siguiente, que tiene la intención de poner énfasis en la diferencia entre ellas.

EJEMPLO Oficinas electas El consejo de fondos de inversión de la universidad a la que asistió el autor se integra con nueve miembros. Cada año, ellos eligen un comité de tres personas para supervisar los edificios y los terrenos. También cada año eligen un presidente, un vicepresidente y un secretario.

a. Cuando el consejo elige el comité de edificios y terrenos, ¿cuántos distintos comités de tres personas son posibles?

b. Cuando el consejo elige a los tres funcionarios (presidente, vicepresidente y secretario), ¿cuántas diferentes planillas de candidatos son posibles?

SOLUCIÓN Note que el orden es irrelevante cuando se elige el comité de edificios y terrenos. Sin embargo, cuando se elige a los funcionarios, los diferentes acomodos cuentan por separado.

a. Puesto que el orden no cuenta para los comités, queremos el número de combinaciones de $r = 3$ personas que se seleccionarán de entre las $n = 9$ disponibles. Tenemos

$$_nC_r = \frac{n!}{(n-r)!r!} = \frac{9!}{(9-3)!3!} = \frac{362{,}880}{4320} = 84$$

continúa

b. Ya que el orden sí cuenta en las planillas de candidatos, queremos el número de secuencias (o permutaciones) de $r = 3$ personas que se seleccionarán de entre las $r = 9$ disponibles. Tenemos

$$_nP_r = \frac{n!}{(n-r)!} = \frac{9!}{(9-3)!} = \frac{362{,}880}{720} = 504$$

Hay 84 diferentes comités de tres miembros del consejo posibles, pero 504 diferentes planillas de candidatos posibles.

Las técnicas de conteo que se presentan en esta sección se usan algunas veces en problemas de probabilidad. Los ejemplos siguientes ilustran dichas aplicaciones.

EJEMPLO Lotería de Maine En la lotería de Maine se extraen seis diferentes números del 1 al 42. Un jugador gana o comparte el premio mayor cuando escoge la combinación correcta de seis números. Si un jugador elige una combinación particular de seis números, calcule la probabilidad de ganar el premio mayor. (No se requiere que el jugador seleccione los seis números en el mismo orden en que se sacaron, por lo que el orden es irrelevante).

SOLUCIÓN Puesto que se seleccionan seis números diferentes de 42 posibilidades diferentes, el número total de combinaciones es

$$_{42}C_6 = \frac{42!}{(42-6)!6!} = \frac{42!}{36!6!} = 5{,}245{,}786$$

Si el jugador sólo selecciona una combinación, la probabilidad de ganar es de $1/5{,}245{,}786$.

EJEMPLO La lotería Powerball La lotería Powerball se juega en 21 estados de Estados Unidos. Usted debe seleccionar cinco números entre 1 y 49, además de otro número especial Powerball entre 1 y 42. (Se extraen cinco bolas de una tómbola con 49 bolas blancas y una bola roja de una tómbola con 42 bolas rojas). El número especial Powerball puede ser el mismo que uno de los otros cinco números. Para ganar o compartir el premio mayor, tendrá que seleccionar la combinación correcta de cinco números *y* también debe seleccionar el número Powerball correcto. Calcule la probabilidad de ganar o compartir el premio mayor.

SOLUCIÓN Dividamos este problema en tres partes: **1.** tener la combinación correcta de cinco números; **2.** obtener el número Powerball correcto, y **3.** combinar los resultados para calcular la probabilidad de ganar o compartir el premio mayor. Comencemos con el número de combinaciones que son posibles cuando usted selecciona cinco números entre 1 y 49, que es

$$_{49}C_5 = \frac{49!}{(49-5)!5!} = \frac{49!}{44!5!} = 1{,}906{,}884$$

La probabilidad de obtener la combinación ganadora de cinco números es, por lo tanto, de $1/1{,}906{,}884$.

continúa

En segundo lugar, seleccionemos también el número Powerball correcto entre 1 y 42. La probabilidad de seleccionar el número Powerball ganador es de $1/42$.

En tercer lugar, puesto que debemos obtener la combinación correcta de cinco números y el número Powerball correcto, la probabilidad de la ocurrencia de ambos sucesos es de $1/1{,}906{,}884 \times 1/42 = 1/80{,}089{,}128$. Este último resultado es una aplicación de la regla de la multiplicación que se explica en la sección 3-4. Para una persona que compre un boleto, la probabilidad de ganar la lotería Powerball es de $1/80{,}089{,}128$.

3-7 Destrezas y conceptos básicos

Cálculo de factoriales, combinaciones y permutaciones. *En los ejercicios 1 a 8 evalúe las expresiones dadas y exprese todos los resultados utilizando el formato normal para escribir números (en lugar de notación científica).*

1. 6! **2.** 15! **3.** ${}_{25}P_2$ **4.** ${}_{100}P_3$

5. ${}_{25}C_2$ **6.** ${}_{100}C_3$ **7.** ${}_{52}C_5$ **8 .** ${}_{52}P_5$

Probabilidad de ganar en la lotería. *Esta sección incluyó un ejemplo que mostró que la probabilidad de ganar la lotería de Maine es de $1/5{,}245{,}786$. En los ejercicios 9 a 12 calcule la probabilidad de ganar la lotería que se indica.*

9. Massachusetts Mass Millions: Seleccionar los seis números ganadores entre 1, 2, . . . , 49.

10. Pennsylvania Super 6 Lotto: Seleccionar los seis números ganadores entre 1, 2, . . . ,69.

11. New York Lotto: Seleccionar los seis números ganadores entre 1, 2, . . . , 59.

12. New York Take Five: Seleccionar los cinco números ganadores entre 1, 2, . . . , 39.

13. Discriminación por edad La empresa Pitt Software Company redujo su personal de ventas de 32 a 28 empleados. La compañía afirmó que seleccionó a cuatro empleados al azar para despedirlos. Sin embargo, los cuatro empleados que eligió son los más viejos de la fuerza de ventas original de 32. Calcule la probabilidad de que cuando se seleccionan cuatro empleados al azar de un grupo de 32, éstos sean los cuatro más viejos. ¿Es la probabilidad lo suficientemente baja como para acusar a la Pitt Software Company de que en lugar de usar selección aleatoria, en realidad sólo despidió a los empleados más viejos?

14. Diseño de computadoras En el diseño de una computadora, un *byte* se define como una secuencia de 8 bits y cada bit debe ser un 0 o un 1. ¿Cuántos bytes diferentes son posibles? (Con frecuencia se usa un byte para representar un carácter individual, como una letra, un dígito o un símbolo de puntuación. Por ejemplo, cierto sistema de codificación representa la letra *A* como 01000001). ¿Existen suficientes bytes diferentes para los caracteres que usamos comúnmente, incluyendo letras minúsculas, letras mayúsculas, dígitos, símbolos de puntuación, signo de pesos y otros?

15. Lotería de Maine La probabilidad de ganar la lotería de Maine es de $1/5{,}245{,}786$. Si las reglas se modificaran para que, además de seleccionar los seis números correctos del 1 a 42, ahora tuviera que elegirlos en el mismo orden en que son extraídos, ¿cuál es la probabilidad de ganar?

16. Prueba de una afirmación Mike afirma que desarrolló la habilidad de obtener un 6 casi siempre que tira un dado. Usted prueba su afirmación haciendo que Mike tire un dado cinco veces, en tanto él obtiene el 6 cada vez. Si Mike no tiene posibilidad de afectar los resultados, calcule la probabilidad de que él tire cinco veces consecutivas y obtenga 6 en todas. ¿Es la probabilidad suficientemente baja como para apoyar la afirmación de Mike?

17. Selección de grupo de tratamiento La empresa Walton Pharmaceuticals quiere probar la eficacia de un nuevo fármaco que se desarrolló para aliviar síntomas de alergia. La prueba inicial se realizará tratando a seis personas que se escogieron de un grupo de 15 voluntarios. Si el grupo de tratamiento se selecciona aleatoriamente, ¿cuál es la probabilidad de que se conforme con las seis personas más jóvenes del grupo? Si se selecciona a los seis más jóvenes, ¿hay evidencia suficiente para concluir que la selección se basó en la edad, en lugar de ser aleatoria?

18. Lo hizo a su manera El legendario cantante Frank Sinatra grabó 381 canciones. Usted debe seleccionar, de una lista de sus 10 más grandes éxitos, tres que serán cantados en un popurrí como un tributo en la próxima ceremonia de los premios MTV Music Awards. El orden de las canciones es importante, ya que tienen que sonar bien juntas. Si usted selecciona tres de las canciones de Sinatra de entre sus 10 mayores éxitos, ¿cuántas secuencias diferentes son posibles?

19. Rutas aéreas Usted acaba de inaugurar su propia empresa de líneas aéreas llamada Air América, cuyo lema es: "Donde su probabilidad de un vuelo seguro es mayor que cero". Trazó un plan para una ruta que conecta Austin, Boise y Chicago. Una ruta es Austin-Boise-Chicago; una segunda ruta es Chicago-Boise-Austin. ¿Cuántas otras rutas son posibles? ¿Cuántas rutas diferentes son posibles si el servicio se expandiera para incluir un total de ocho ciudades?

20. Números del Seguro Social Cada número del Seguro Social es una secuencia de nueve dígitos. ¿Cuál es la probabilidad de generar aleatoriamente nueve dígitos y obtener su número de Seguro Social?

21. Electrizante Para probar la corriente eléctrica en un conductor con cables codificados en cinco colores, el autor utilizó un medidor para probar dos cables a la vez. ¿Cuántas pruebas se requieren para verificar cada posible par de cables?

22. Consejo de administración electo En un consejo de administración del hospital general de Newport hay 12 miembros.

a. Si ellos deben elegir un presidente, un primer vicepresidente, un segundo vicepresidente y un secretario, ¿cuántas planillas de candidatos diferentes son posibles?

b. Si tienen que formar un subcomité de ética de cuatro miembros, ¿cuántos subcomités diferentes son posibles?

23. Sopa de letras Muchos periódicos incluyen una "sopa de letras", un crucigrama donde el lector debe descifrar letras para formar palabras. Por ejemplo, las letras TAISER se incluyeron en un periódico del día en que se escribió este ejercicio. ¿De cuántas formas se pueden acomodar las letras TAISER? Identifique la palabra que se codificó y luego determine la probabilidad de obtener este resultado seleccionando al azar un arreglo de las letras dadas.

24. Calcular el número de melodías posibles En el *Directorio de melodías y temas musicales* de Dennys Parsons, se listan melodías de más de 14,000 canciones de acuerdo con el siguiente esquema: la primera nota de cada canción se representa con un asterisco (*), en tanto que las notas sucesivas lo hacen con una *R* (para repetir la nota previa), *S* (para una nota que sube) o *B* (para una nota que baja). La quinta sinfonía de Beethoven comienza como *RRB. Se representan melodías clásicas mediante las primeras 16 notas. Con este esquema, ¿cuántas melodías clásicas diferentes son posibles?

25. Candados de combinación Un candado "de combinación" común se abre con la secuencia correcta de tres números entre 0 y 49, inclusive. (Es posible utilizar un número más de una vez). ¿Cuál es la probabilidad de adivinar los tres números y de abrir el candado en el primer intento?

26. Flor de cinco naipes Un mazo de naipes normal contiene 13 tréboles, 13 diamantes, 13 corazones y 13 espadas. Si se eligen cinco naipes aleatoriamente, calcule la probabilidad de obtener una flor. (Se tiene una flor cuando las cinco cartas son del mismo palo. Es decir, cuando todas son tréboles, diamantes, corazones o espadas).

27. Probabilidades de secuencias de género

a. Si una pareja planea tener ocho hijos, ¿cuántas secuencias de género diferentes son posibles?

b. Si una pareja tiene cuatro niños y cuatro niñas, ¿cuántas secuencias de género diferentes son posibles?

c. Con base en los resultados de los incisos *a* y *b*, cuando una pareja tiene ocho hijos ¿cuál es la probabilidad de que sean cuatro niños y cuatro niñas?

28. ¿El investigador está haciendo trampa? Cuando un investigador de genética selecciona al azar grupos de 20 bebés recién nacidos y aparentemente obtiene 10 niñas y 10 niños con consistencia, usted se vuelve suspicaz. El investigador explica que es común obtener 10 niños y 10 niñas en estos casos.

a. Si se seleccionan 20 bebés recién nacidos, ¿cuántas secuencias de género distintas son posibles?

b. ¿De cuántas formas diferentes pueden acomodarse, en secuencia, 10 niños y 10 niñas?

c. ¿Cuál es la probabilidad de obtener 10 niños y 10 niñas cuando nacen 20 bebés?

d. Con base en los resultados de lo anterior, ¿está de acuerdo con la explicación del investigador de que es común obtener 10 niños y 10 niñas cuando se seleccionan 20 bebés al azar?

29. Calcular el número de códigos de área El reportero Paul Wiseman, del diario *USA Today*, describió las viejas reglas para los códigos de área telefónicos al escribir acerca de "códigos de área posibles con 1 o 0 en el segundo dígito". (Excluidos: códigos que terminen en 00 y 11, para llamadas con cargo gratuito, servicios de emergencia y otros usos especiales). Los códigos que empiezan con 0 o 1 también deben ser excluidos. ¿Cuántos códigos de área distintos era posible obtener bajo tales viejas reglas?

30. Huevos rotos Una caja contiene 12 huevos, tres de los cuales están rotos. Si seleccionamos al azar cinco de los huevos para cocerlos, ¿cuál es la probabilidad de los sucesos siguientes?:

a. Todos los huevos que se seleccionaron están rotos.

b. Ninguno de los huevos que se seleccionaron está roto.

c. Dos de los huevos que se seleccionaron están rotos.

31. Lotería de California En el juego de lotería Super Lotto Plus de California, ganar el premio mayor requiere que se seleccionen los cinco números correctos del 1 al 47 y, por separado, también elegir un solo número correcto entre 1 y 27. Calcule la probabilidad de ganar el premio mayor.

32. Torneo de basquetbol NCAA Cada año, 64 equipos universitarios de basquetbol compiten en el torneo de la NCAA. Recientemente, Sandbox.com ofreció un premio de 10 millones de dólares a cualquiera que eligiera al ganador en todos y cada uno de los juegos del torneo. El presidente de esa compañía también prometió que, además de entregar el premio en efectivo, se comería los gusanos que cupieran en una cubeta. ¡Qué asco!

a. ¿Cuántos juegos se requieren para obtener un equipo campeón en un campo de 64 equipos?

b. Si alguna persona hace conjeturas al azar para cada juego del torneo, calcule la probabilidad de escoger al ganador de cada juego.

c. En un artículo acerca del premio de 10 millones de dólares, el diario *New York Times* escribió que "aun un experto en basquetbol colegial que pudiera escoger juegos con acierto en una porción del 70% tiene una probabilidad de 1 en _______ de elegir todos los juegos acertadamente". (Llene el espacio).

3-7 Más allá de lo básico

33. Cálculo del número de nombres de variables de cómputo Una regla común de programación de computadoras es que los nombres de las variables deben tener una longitud de 1 a 8 caracteres. El primer carácter puede ser cualquiera de las 26 letras, mientras que los caracteres sucesivos serían cualesquiera de las 26 letras o de los 10 dígitos. Por ejemplo, A, BBB y M3477K son nombres permitidos de variables. ¿Cuántos nombres de variables diferentes son posibles?

34. Saludos y mesas redondas

a. Cinco gerentes se reúnen para una junta. Si cada uno saluda estrechando la mano a los otros gerentes exactamente una vez, ¿cuál es el número total de saludos?

b. Si *n* gerentes se saludan con cada uno de los otros exactamente una vez, ¿cuál es el número total de saludos?

c. ¿De cuántas maneras diferentes se pueden sentar cinco gerentes en una mesa redonda? (Suponga que si cada uno se mueve a la derecha el acomodo es el mismo).

d. ¿De cuántas formas diferentes se pueden sentar *n* gerentes en una mesa redonda?

35. Evaluación de factoriales grandes Muchas calculadoras o computadoras no pueden calcular directamente el número 70! o un número factorial mayor. Cuando *n* es grande, $n!$ puede aproximarse a $n = 10^{K}$, donde $K = (n + 0.5)\log n + 0.39908993 - 0.43429448n$.

a. Evalúe 50! usando la tecla factorial de una calculadora y también la aproximación que damos aquí.

a. El Departamento de Pesca una vez pidió ayuda a los Laboratorios Bell con la finalidad de encontrar la ruta más corta para obtener muestras en 300 emplazamientos del Golfo de México. Si usted calcula el número de posibles rutas diferentes, ¿cuántos dígitos se necesitan para escribir el número?

36. Inteligencia artificial ¿Las computadoras "piensan"? De acuerdo con la *prueba Turing*, se considera que una computadora piensa, cuando alguien se comunica con ella, si la persona que la utiliza cree que se está comunicando con otro y no con una máquina. En un experimento en el Computer Museum de Boston, cada uno de 10 jueces se comunicó con cuatro de estas máquinas y otras cuatro personas; se les pidió que distinguieran entre ellos.

a. Suponga que el primer juez no puede distinguir entre las cuatro computadoras y las cuatro personas. Si este juez hace conjeturas al azar, ¿cuál es la probabilidad de identificar correctamente las cuatro computadoras y las cuatro personas?

b. Suponga que ninguno de los 10 jueces puede distinguir entre las computadoras y las personas, por lo que hacen conjeturas al azar. Con base en el resultado del inciso *a*, ¿cuál es la probabilidad de que los 10 jueces acierten en todas sus conjeturas? (Este suceso nos permitiría concluir que las computadoras no pueden "pensar" cuando, de acuerdo con la prueba Turing, sí es así).

Repaso

Iniciamos este capítulo con el concepto básico de probabilidad, el cual es de suma importancia para los métodos de estadística inferencial que se introducen más adelante. Aprendimos que un valor de probabilidad, que se expresa como un número entre 0 y 1, refleja la posibilidad de ocurrencia de algún suceso. También que un valor como 0.01 representa un suceso que tiene muy pocas posibilidades de ocurrir. En la sección 3-1, introdujimos la regla del suceso infrecuente para estadística inferencial: si, bajo un supuesto dado, la probabilidad de un suceso particular es extremadamente pequeña, concluimos que quizás el supuesto es incorrecto. Como ejemplo del enfoque básico que se utilizó, considere la prueba de la aseveración de alguien de que una moneda que se usa en un volado está balanceada. Si lanzamos la moneda 10 veces y obtenemos 10 caras consecutivas, de estos resultados de muestra es posible hacer una de dos inferencias:

1. La moneda está realmente balanceada y la cadena de 10 caras consecutivas es una *chiripa*.
2. La moneda no está balanceada.

Los estadísticos usan la regla del suceso infrecuente cuando deciden cuál inferencia es correcta: en este caso, la probabilidad de obtener 10 caras consecutivas es tan pequeña (1/1024) que la inferencia de que la moneda no está balanceada es la mejor opción. Aquí observamos el importante papel de la probabilidad en los métodos estándar de inferencia estadística.

En la sección 3-2 presentamos definiciones y notaciones básicas, incluyendo la representación de sucesos por letras como A. Definimos, asimismo, las probabilidades de un suceso simple como

$$P(A) = \frac{\text{número de veces que ocurre } A}{\text{número de veces que se repite el experimento}} \qquad \text{(frecuencia relativa)}$$

$$P(A) = \frac{\text{número de formas en que puede ocurrir } A}{\text{número de sucesos simples diferentes}} = \frac{s}{n} \qquad \text{(para resultados igualmente probables)}$$

Además, señalamos que la probabilidad de cualquier suceso imposible es 0 y la probabilidad de cualquier suceso inevitable es 1, así como que para cualquier suceso A, $0 \leq P(A) \leq 1$. También, que $\overline{A}$ denota el complemento del suceso A, es decir, indica que el suceso A no ocurre.

En las secciones 3-3, 3-4 y 3-5 consideramos sucesos compuestos, los cuales combinan dos o más sucesos simples. Asociamos "o" con la suma y asociamos "y" con la multiplicación. Siempre tome en cuenta las consideraciones clave siguientes:

- Cuando se realiza un ensayo, ¿queremos la probabilidad del suceso A o B? Si es así, use la regla de la suma, pero sea cuidadoso a fin de evitar contar cualquier resultado más de una vez.
- Cuando se busca la probabilidad de que el suceso A ocurra en un ensayo *y* el suceso B ocurra en un segundo ensayo, use la regla de la multiplicación. Multiplique la probabilidad del suceso A por la probabilidad del suceso B. *Precaución:* Cuando calcule la probabilidad del suceso B, asegúrese de tomar en cuenta el hecho de que el suceso A ocurrió realmente.

En algunos problemas de probabilidad, el mayor obstáculo es encontrar el número total de resultados posibles. La sección 3-7 se dedicó a las siguientes técnicas de conteo:

- Regla de conteo fundamental
- Regla factorial
- Regla de las permutaciones (cuando todos los elementos son diferentes)
- Regla de las permutaciones (cuando algunos elementos son iguales a otros)
- Regla de las combinaciones

Ejercicios de repaso

Detectores de mentiras. *En los ejercicios 1 a 8 use los datos de la tabla adjunta (que se basa en datos de la Office of Technology Assessment). Los datos reflejan las respuestas a una pregunta clave que se hizo a 100 sujetos diferentes.*

	El polígrafo indicó verdad	El polígrafo indicó mentira
El sujeto realmente dijo la verdad	65	15
El sujeto realmente mintió	3	17

1. Si se selecciona al azar a 1 de los 100 sujetos, calcule la probabilidad de obtener a alguien que mintió.

2. Si se selecciona al azar a 1 de los 100 sujetos, calcule la probabilidad de obtener a alguien para quien la prueba del polígrafo indicó que dijo una mentira.

3. Si se selecciona al azar a 1 de los 100 sujetos, calcule la probabilidad de obtener a alguien que mintió o tuvo la indicación de la prueba del polígrafo de que lo hizo.

4. Si se selecciona al azar a 1 de los 100 sujetos, calcule la probabilidad de obtener a alguien que dijo la verdad o la prueba del polígrafo indicó que respondió con la verdad.

5. Si dos diferentes sujetos se seleccionan al azar, calcule la probabilidad de que ambos dijeron la verdad.

6. Si dos diferentes sujetos se seleccionan al azar, calcule la probabilidad de que, al realizar la prueba del polígrafo, éste indicó que ambos dijeron una mentira.

7. Si un sujeto se selecciona al azar, calcule la probabilidad de que él o ella dijeron la verdad, puesto que la prueba del polígrafo indicó que era una mentira.

8. Si un sujeto se selecciona al azar, calcule la probabilidad de que él o ella obtuvieron la indicación de la prueba del polígrafo de que dijeron una mentira, puesto que el sujeto realmente dijo la verdad.

9. Probabilidad de fallas de computadora Una encuesta de la revista *PC World* entre 4,000 personas propietarias de computadoras personales, mostró que 992 de ellas se averiaron durante los primeros dos años (se averiaron las computadoras, no las personas). Al seleccionar entre varios distribuidores de computadoras, un agente de compras quiere saber la probabilidad de que una computadora personal se descomponga durante los primeros dos años. Utilice los resultados de la encuesta para estimar esa probabilidad.
   a. Si se selecciona una computadora personal al azar, ¿cuál es la probabilidad de que se descomponga durante los primeros dos años?
   b. Si se seleccionan dos computadoras personales al azar, ¿cuál es la probabilidad de que ambas se descompongan durante los dos primeros años?
   c. Si se seleccionan tres computadoras personales al azar, ¿cuál es la probabilidad de que al menos una de ellas se descomponga durante los dos primeros años?

10. Muestreo de aceptación Con un método de muestreo de aceptación, una muestra de artículos se selecciona aleatoriamente sin reemplazo y el lote completo se rechaza si se encuentra al menos un defecto. La Niko Electronics Company acaba de fabricar 2,500 CD, de los cuales el 2% salieron defectuosos. Si se seleccionan y prueban cuatro de los CD, ¿cuál es la probabilidad de que se rechace el lote completo?

11. Prueba de una aseveración La Biogene Research Company afirma que desarrolló una técnica para asegurar que un bebé será una niña. En una prueba de esa técnica, 12 parejas tuvieron niñas. Calcule la probabilidad de obtener dos niñas por casualidad, suponiendo que los niños y las niñas son igualmente probables, así como que el género de cualquier hijo es independiente del género de los otros. ¿Apoya el resultado la aseveración de la compañía?

12. Selección de miembros El Consejo de Administración del Hartford Investment Fund cuenta con 10 miembros.
   a. Si se selecciona al azar a tres miembros para supervisar a los auditores, calcule la probabilidad de que sean seleccionados los tres miembros más acaudalados.
   b. Si se eligen miembros para los puestos de presidente, vicepresidente y tesorero, ¿cuántas planillas diferentes son posibles?

13. Ruleta Cuando se apuesta a *pares* en la ruleta, hay 38 resultados igualmente probables, pero sólo 2, 4, 6, . . . , 36 son resultados ganadores.
a. Calcule la probabilidad de ganar cuando se apuesta a pares.
b. Calcule las posibilidades reales en contra de ganar cuando se apuesta a pares.
c. Los casinos pagan las apuestas ganadoras de acuerdo con las posibilidades descritas como 1:1. ¿Cuál sería su ganancia neta si apuesta $5 a pares y gana?

14. ¿Está mintiendo el encuestador? Un encuestador afirma que 12 votantes se seleccionaron aleatoriamente de una población de 200,000 (el 30% de ellos son republicanos) y que los 12 fueron republicanos. El encuestador añade que esto podría suceder fácilmente por casualidad. Calcule la probabilidad de obtener 12 republicanos cuando se seleccionan 12 votantes de dicha población. Con base en el resultado, ¿parece ser correcta la afirmación del encuestador?

15. Seguros de vida La compañía de seguros New England Life expide pólizas anuales a 12 hombres de 27 años de edad. Con base en datos del Department of Health and Human Services, cada uno de estos individuos tiene una posibilidad de un 99.82% de vivir todo el año. ¿Cuál es la probabilidad de que todos ellos sobrevivan el año?

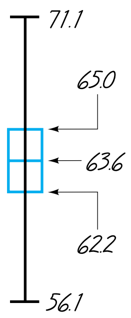
16. Loterías de Illinois Illinois tiene diferentes juegos de lotería. Calcule la probabilidad de ganar el premio mayor en cada juego.
a. Lotto: Seleccionar los seis números ganadores de 1, 2, . . . , 52.
b. Little Lotto: Seleccionar los cinco números ganadores entre 1, 2, . . . , 30.
c. The Big Game: Seleccionar los cinco números ganadores entre 1, 2, . . . , 50 *y*, además, seleccionar también un solo número ganador entre 1, 2, . . . , 36.

Ejercicios de repaso acumulativos

1. Tratamiento del síndrome de fatiga crónica A una muestra de pacientes que padecen el síndrome de fatiga crónica se le trató con medicamentos, después se midió el cambio en su fatiga en una escala de -7 a $+7$, con los valores positivos representando mejoría y con 0 representando ningún cambio. Los resultados se listan abajo (con base en datos de "The Relation Between Neurally Mediated Hypotension and the Chronic Fatigue Syndrome", de Bou-Holaigah, Rowe, Kan y Calkins, *Journal of the American Medical Association*, vol. 274, núm. 12).

6 5 0 5 6 7 3 3 2 4 4 0 7 3 4 3 6 0 5 5 6

a. Calcule la media.
b. Calcule la mediana.
c. Calcule la desviación estándar.
d. Calcule la varianza.
e. Con base en los resultados, ¿fue efectivo el tratamiento?
f. Si se selecciona aleatoriamente un valor de esta muestra, calcule la probabilidad de que sea positivo.
g. Si se seleccionan aleatoriamente dos diferentes valores de esta muestra, calcule la probabilidad de que ambos sean positivos.
h. Ignore los tres valores de 0 y suponga que sólo son posibles valores positivos o negativos. Suponiendo que el tratamiento no es efectivo y que los valores positivos y negativos son igualmente probables, calcule la probabilidad de que 18 sujetos tengan todos valores positivos (como en este grupo muestral). ¿Es esta probabilidad bastante baja como para justificar el rechazo de la suposición de que el tratamiento no es efectivo?



2. Estaturas de mujeres La gráfica de cuadro anexa describe estaturas (en pulgadas) de una gran colección de mujeres adultas que se seleccionaron aleatoriamente.
   a. ¿Cuál es el promedio de estaturas de las mujeres adultas?
   b. Si se selecciona al azar a una de estas mujeres, calcule la probabilidad de que su estatura se ubique entre 56.1 pulgadas y 62.2 pulgadas.
   c. Si se selecciona al azar a una de estas mujeres, calcule la probabilidad de que su estatura sea más baja que 62.2 pulgadas o que sobrepase 63.6 pulgadas.
   d. Si se seleccionan dos mujeres al azar, calcule la probabilidad de que ambas tengan estaturas entre 62.2 pulgadas y 63.6 pulgadas.
   e. Si se seleccionan cinco mujeres al azar, calcule la probabilidad de que tres de ellas sean más altas que la media y las otras dos sean más bajas que la media.

Actividades de cooperación en equipo

1. Actividad en clase Véase el ejercicio 15 de la sección 3-6. Formen equipos de tres o cuatro y realicen lanzamientos de monedas para desarrollar una simulación que imite al reino que se atiene a tal decreto: después de que una madre dé a luz a un hombre, ella no tendrá ningún otro hijo. Si este decreto se obedece, ¿se incrementará la proporción de mujeres?

2. Actividad en clase Haga equipos de tres o cuatro personas y use tachuelas para estimar la probabilidad de que, cuando se dejan caer, una tachuela quede con la punta hacia arriba. ¿Cuántos intentos son necesarios para obtener un resultado que parezca razonablemente preciso, cuando se redondea al primer espacio decimal?

3. Actividad fuera de clase Los biólogos marinos con frecuencia usan el método captura-recaptura como procedimiento para estimar el tamaño de una población como, por ejemplo, del número de peces en un lago. Este método consiste en capturar una muestra de la población, etiquetar a cada uno de los miembros de la muestra y luego regresarlos a la población.

   En lugar de capturar peces reales, simule el procedimiento utilizando un conjunto uniforme de artículos como, por ejemplo, botones, cuentas de colores, dulces M&M, piezas de cereal de aros de frutas o tarjetas de archivo. Comience con una colección grande de dichos artículos. Obtenga una muestra de 50 y use un marcador para "etiquetar" a cada uno. Reemplace los artículos que se marcaron, revuelva la población completa; luego, seleccione una segunda muestra y proceda a estimar el tamaño de la población. Compare el resultado con el tamaño real de la población que se obtiene contando todos los artículos.

4. Actividad en clase Formen equipos de dos. Remítanse al ejercicio 13 en la sección 3-6 para tener una descripción del "problema Monty Hall". Simulen el concurso, y registren los resultados de quedarse y de cambiar; después, determinen cuál de estas dos estrategias es mejor.

5. Actividad fuera de clase Formen equipos de dos, con el propósito de hacer un experimento que se diseñó con la finalidad de mostrar un enfoque para el manejo de preguntas de encuesta sensibles, que se relacionan con el uso de drogas, la actividad sexual (o la inactividad), el robo o la estafa. En vez de utilizar realmente una pregunta polémica que podría ocasionar ira contra el autor, manejemos esta inocua pregunta: "¿Nació usted en un mes que tiene la letra *r*?". Alrededor de 2/3 de todas las respuestas deben ser "sí", pero vamos a suponer que la pregunta es muy sensible y que esos sujetos de la encuesta son reticentes a contestar con honestidad. Encueste pidiéndoles a las personas que lancen una moneda al aire y respondan como sigue:

   - "Sí", si la moneda cae en cruz *o* usted nació en un mes que tiene la letra *r*.
   - "No", si la moneda cae en cara *y* usted nació en un mes que no contiene la letra *r*.

   Supuestamente, quienes responden tienden a ser más honestas, porque sienten que lanzar la moneda al aire protege su privacidad. Encueste personas y analice los resultados para determinar la proporción de aquellos que nacieron en un mes que contiene la letra *r*. La precisión de los resultados se puede cotejar con las fechas de nacimiento reales, que se obtendrán de una segunda pregunta. Es posible repetir el experimento con una pregunta que sea más sensible, pero aquí no se plantean preguntas de este tipo, porque el autor ya recibe suficientes mensajes por correo.

Proyecto tecnológico

Este proyecto ilustra la ley de los números grandes descrita en la sección 3-2. Use una computadora o la calculadora TI-83 Plus para simular 100 nacimientos. Realice esto generando aleatoriamente 100 números, ya sea 0 o 1 (donde 0 = niño y 1 = niña). Utilice los resultados para completar la tabla siguiente. ¿Qué sucede a la proporción de niñas conforme el tamaño de la muestra se va incrementando? ¿Cómo ilustra esto la ley de los números grandes?

Número de nacimientos	10	20	30	40	50	60	70	80	90	100
Proporción de niñas										

Aquí tenemos algunos detalles para las diferentes aplicaciones tecnológicas.

STATDISK Seleccione **Data** en la barra de menú principal. Elija **Uniform Generator**. Realice las entradas para un tamaño de muestra de 100, un mínimo de 0, un máximo de 1 y 0 espacios decimales (puesto que queremos números enteros).

Minitab Seleccione **Calc** de la barra de menú principal en la parte superior. Seleccione **Random Data**; después, **Integer**. Proceda a cargar 100 para el número de filas de datos, C1 para la columna en la que se guardarán los resultados, 0 para el valor mínimo y 1 para el valor máximo. Cuando termine, haga clic en **OK**.

Excel Posicione el cursor en la celda A1. Haga clic en el elemento del menú f_x; luego, seleccione **Math & Trig** y **RANDBETWEEN**. Haga clic en **OK**. En el cuadro de diálogo, ponga 0 para el valor inferior y 1 para el valor superior; luego, haga clic en **OK**. La celda A1 debe contener ahora un 0 o un 1. Haga clic en la esquina inferior derecha de esa celda y, mientras sostiene el botón del ratón, arrastre el cursor hacia abajo hasta la celda A100, ahora suéltelo.

TI-83 Plus **Plus** Oprima la tecla **MATH**. Seleccione **PRB**. Luego, el 5o elemento del menú: **randInt**. Ingrese 0, 1, 100, y presione la tecla **ENTER**. Oprima **STO** y **L1** para guardar los datos en la lista L1. Para ver los nacimientos que se generaron, oprima **STAT** y seleccione **Edit**.

de los DATOS *a la* DECISIÓN

Pensamiento crítico: Cuando usted solicita un empleo, ¿debe preocuparse por la prueba de consumo de drogas?

De acuerdo con la American Management Association, alrededor del 70% de las empresas de Estados Unidos ya realizan la prueba del consumo de drogas al menos a algunos empleados y solicitantes de empleo. El National Insitute on Drug Abuse afirma que alrededor del 15% de las personas entre 18 y 25 años consumen drogas ilegales. Allyn Clark, un universitario graduado de 21 años de edad, cuando solicitó empleo en la compañía Acton Paper, se sometió a la prueba de drogas; en consecuencia, no se le dio el empleo. Él sospechó que tal vez obtuvo un resultado adverso en la prueba de drogas, aunque no las consume. Revisando con el departamento de personal de la empresa, encontró que sólo el 1% de los usuarios de drogas arrojan erróneamente un resultado de prueba negativo y que sólo el 2% de quienes no consumen drogas se identifican incorrectamente como usuarios de drogas. Allyn se sintió aliviado por estas cifras, porque creyó que reflejaban una prueba muy confiable que generalmente daba buenos resultados, pero, ¿es esto realmente cierto?

Análisis de los resultados

La tabla adjunta muestra datos de Allyn y de otros 1999 solicitantes de empleo. Con base en estos resultados, calcule *P*(falso negativo); esto es, la probabilidad de seleccionar al azar a alguna persona que tuvo un resultado de prueba positivo y elegir a alguien que no consume drogas. ¿Son las probabilidades de dichos resultados erróneos suficientemente bajas como para que los solicitantes de empleo y la compañía Acton Paper no tengan de qué preocuparse?

	Usuarios de drogas	No usuarios
Resultado de prueba positivo	297	34
Resultado de prueba negativo	3	1666

PROYECTO DE INTERNET

Proceso de probabilidades por computadora

Calcular las probabilidades cuando se tiran dados es fácil. Con un dado, hay seis posibles resultados, de los cuales cada uno, como por ejemplo tirar un 2, tiene una probabilidad de $1/6$. Para un juego de naipes se necesitan más cálculos, aunque siguen siendo manejables. Pero, ¿qué pasa con un juego más complicado, como por ejemplo el juego de mesa Monopolio? ¿Cuál es la probabilidad de aterrizar en un lugar en particular del tablero? La probabilidad depende del lugar que su pieza ocupe en el momento del resultado de los dados, de tomar cartas y de otros factores. Ahora considere un ejemplo más representativo de la vida real, como el de la probabilidad de tener un accidente automovilístico. El número de factores implicados es muy grande como para siquiera considerarlos; no obstante, las compañías de seguros, por ejemplo, contemplan probabilidades de este tipo.

El proyecto de Internet para este capítulo considera métodos para calcular probabilidades en situaciones que se complican. Vaya al proyecto de Internet, que encontrará en este sitio:

http://www.pearsoneducacion.net/triola

Se le guiará en la investigación de probabilidades para un juego de mesa. Después, calcule usted mismo probabilidades de este tipo. Finalmente, efectuará un estimado de una probabilidad que se relaciona con la salud utilizando datos empíricos.

La estadística @ en el trabajo

"Con base en estas estadísticas acerca del uso preferido del parque, evaluamos cómo dar mejor servicio a la población más diversa".

Judy Shafer

Segunda superintendente del Parque Nacional Virgin Islands

Judy trabaja para el National Park Service desde hace 17 años y para la administración del parque desde los últimos cuatro.

Como segunda superintendente del Parque Nacional Virgin Islands, ¿utiliza la estadística en su trabajo?

Usamos probabilidad y métodos de estadística en aplicaciones tales como el análisis de la sustentabilidad de una especie particular de coral, bajo las presiones ambientales del uso del parque por el visitante, la contaminación, la sedimentación, etcétera. También, métodos de estadística para determinar cuáles poblaciones étnicas y raciales utilizan el parque, y la manera en la que el parque está logrando sus metas de satisfacción al usuario.

¡Hay tantos usos de la estadística que es imposible describirlos todos! Nuestros guardabosques recolectan datos y usan la estadística para actividades de visita y de protección de los recursos. Nuestros científicos de investigación marina utilizan la estadística a diario para determinar la salud y viabilidad de varias especies marinas, así como para conocer el nivel de amenaza que representan los agentes ambientales (como la contaminación y la sedimentación) y los visitantes del parque. Por ejemplo, algunas de las especies de peces más grandes, como el *Nassau Grouper*, pueden mermarse tanto por la pesca excesiva que tal vez estén cerca de la extinción. Los datos estadísticos de las investigaciones deben guiar las decisiones del parque acerca de cómo proteger dichas especies, ya que se encuentran inextricablemente encadenadas al ecosistema de los grandes arrecifes de coral e incluso al calentamiento global.

¿Podría dar un ejemplo simple y específico de cómo se usa la estadística?

En lo que respecta al uso del parque por el visitante, el National Park Service ha aprendido que distintos grupos étnicos y raciales utilizan los parques nacionales en diferentes formas. Por ejemplo, descubrimos que a algunos grupos les gusta la tradicional caminata campirana mientras que otros prefieren la actividad social de los días de campo en compañía. Estamos considerando la incorporación de más áreas para días de campo en compañía con la finalidad de acomodar mejor a los visitantes de diferentes contextos étnicos y raciales.

¿Es necesario que quienes solicitan empleo en su área tengan un curso de estadística en su historial?

Nuestros empleados deben tener al menos un curso de estadística de nivel universitario. Sin esto, serás un dinosaurio en un siglo XXI que se basa en la estadística. Sin la estadística, no serás tan competitivo como la persona que sí recibió esa preparación.

¿Qué otras habilidades es importante que los empleados posean?

Visión. Liderazgo. Innovación. Creatividad. Y la disposición de tomar algunos riesgos para lograr sus metas y proteger las cosas por las que sienten pasión.

4

Distribuciones de probabilidad

PROBLEMA 4 DEL CAPÍTULO

Determinar si un método de selección del género es efectivo

Los avances recientes en medicina, genética y tecnología son sorprendentes. Ya parece posible la clonación de seres humanos. Las operaciones de puenteo arterial cardiaco, hasta hace poco consideradas riesgosas y peligrosas, ahora son rutina. La cirugía con láser es capaz de corregir la visión, de modo que muchas personas pueden deshacerse de sus anteojos o lentes de contacto. En lugar de confiar sólo en la probabilidad aleatoria, las parejas que planean tener bebés disponen de técnicas para determinar el género de sus hijos. En ocasiones, dichos avances van acompañados de una gran polémica. Algunos consideran que las técnicas de clonación o de selección del género conllevan graves implicaciones morales y que deben prohibirse estrictamente, sin importar cuál sea su justificación. Lisa Belkin escribió lo siguiente en el artículo "Getting the girl" (del *New York Times Magazine):* "Si permitimos que los padres elijan el sexo de sus hijos, ¿cuánto tiempo pasará antes de que ordenen el color de ojos y de cabello, rasgos de personalidad y CI? Hay algunos argumentos convincentes que están a favor del uso, aunque sea limitado, de la selección del género. Uno de esos argumentos señala que ciertas parejas son portadoras de genes recesivos que se relacionan con el cromosoma X, lo cual implica que cualquier hijo hombre cuenta con un 50% de probabilidades de heredar una enfermedad grave, pero que ninguna niña heredará el trastorno. Dichas parejas podrían utilizar la selección del género como una forma para asegurarse de que tendrán niñas y de que ninguna heredará un padecimiento grave.

El Genetics and IVF Institute, de Fairfax, Virginia, creó una técnica llamada MicroSort que, se supone, aumenta las posibilidades de que una pareja tenga una niña. En una prueba preliminar, se reunieron 14 parejas que deseaban niñas. Con el uso de la técnica MicroSort, 13 parejas procrearon niñas y una tuvo un niño. Tales resultados nos conducen a plantearnos una interesante pregunta: puesto que 13 de las 14 parejas tuvieron niñas, ¿concluiríamos que la técnica MicroSort es efectiva o sólo explicaríamos el resultado como la consecuencia de una muestra aleatoria? Para responder lo anterior, usaremos principios de probabilidad que determinen si los nacimientos que se observaron difieren de manera significativa de los resultados que se esperarían al azar. Este ejemplo hace surgir un tema que es esencial para la estadística inferencial: ¿De qué forma determinamos si los resultados deben atribuirse al azar o a un factor como el uso de la técnica MicroSort de selección del género?

4-1 Panorama general

En este capítulo combinamos los métodos de *estadística descriptiva* que se presentan en el capítulo 2 y los de *probabilidad* que se estudiaron en el capítulo 3. La figura 4-1 presenta un resumen esquemático de los objetivos. Como se observa en la figura, con el uso de los métodos del capítulo 2, tiraríamos en repetidas ocasiones un dado para reunir datos muestrales, que luego pueden describirse con gráficas (como un histograma o una gráfica de cuadro), medidas de tendencia central (como la media) y medidas de variación (como la desviación estándar). Con los métodos del capítulo 3, calcularíamos la probabilidad de cada resultado. Ahora combinaremos dichos conceptos mientras creamos distribuciones de probabilidad que describan lo que *probablemente* sucederá, en lugar de lo que en realidad *sucedió*. En el capítulo 2 elaboramos tablas de frecuencias e histogramas con los valores muestrales *observados* que se reunieron en realidad, aquí construiremos distribuciones de probabilidad presentando los resultados posibles, junto con las frecuencias relativas que *esperamos*.

El "empleado" de un casino conoce la forma en que un dado debe comportarse. La tabla en el extremo derecho de la figura 4-1 representa una distribución de probabilidad que sirve como modelo para una distribución de frecuencias poblacional teóricamente perfecta. En esencia, es posible describir la tabla de frecuencias relativas para un dado que se tiró un número infinito de veces. Con tal conocimiento de los resultados de la población, seremos capaces de calcular sus características importantes, tales como la media y la desviación estándar. El resto del libro y la esencia de la estadística inferencial se basan en el conocimiento de las distribuciones de probabilidad. Iniciamos examinando el concepto de una variable aleatoria, después estudiaremos distribuciones importantes con muchas aplicaciones reales.

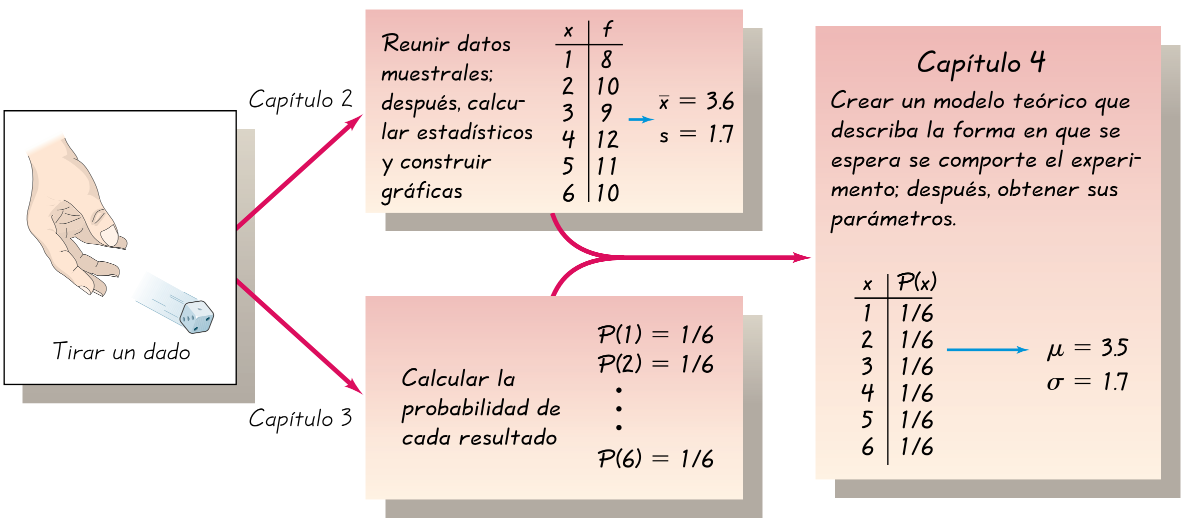


FIGURA 4-1 Combinación de métodos descriptivos y probabilidades para formar un modelo teórico de comportamiento

4-2 Variables aleatorias

En esta sección estudiaremos las variables aleatorias, las distribuciones de probabilidad, los procedimientos para calcular la media y la desviación estándar de una distribución de probabilidad, así como los métodos para distinguir entre resultados que pueden ocurrir por azar y aquellos que son "poco comunes". Iniciaremos con los conceptos que se relacionan de *variable aleatoria* y *distribución de probabilidad*.

Definiciones

Variable aleatoria: variable (casi siempre representada por x) que tiene un solo valor numérico, determinado por el azar, para cada resultado de un procedimiento.

Distribución de probabilidad: gráfica, tabla o fórmula que da la probabilidad de cada valor de la variable aleatoria.

PROBLEMA 4 DEL CAPÍTULO

EJEMPLO Género de niños Un estudio consiste en la selección aleatoria de 14 bebés recién nacidos y el conteo del número de niñas (como en el problema del capítulo). Si consideramos que la probabilidad de niños y niñas es la misma, y

$$x = \text{número de niñas de entre 14 bebés}$$

entonces x es una variable aleatoria, porque su valor depende del azar. Los valores posibles de x son 0, 1, 2, . . . , 14. La tabla 4-1 incluye los valores de x, junto con las probabilidades correspondientes. (En la sección 4-3 aprenderemos a calcular los valores de probabilidad, como los que se listan en la tabla 4-1). Ya que la tabla 4-1 incluye la probabilidad para cada valor de la variable aleatoria x, dicha tabla describe una distribución de probabilidad.

Tabla 4-1
Probabilidades de niñas

x (niñas)	$P(x)$
0	0.000
1	0.001
2	0.006
3	0.022
4	0.061
5	0.122
6	0.183
7	0.209
8	0.183
9	0.122
10	0.061
11	0.022
12	0.006
13	0.001
14	0.000

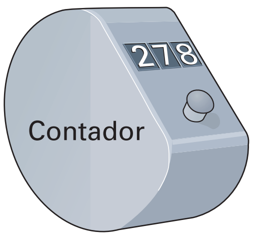
En la sección 1-2 hicimos una distinción entre los datos discretos y los continuos. Las variables aleatorias también pueden ser discretas o continuas, en tanto que las siguientes dos definiciones son consistentes con las que se presentan en la sección 1-2.

Definiciones

Variable aleatoria discreta: tiene un número finito de valores o un número de valores contable, donde "contable" se refiere al hecho de que podría haber un número infinito de valores, pero que pueden asociarse con un proceso de conteo.

Variable aleatoria continua: tiene un número infinito de valores; dichos valores pueden asociarse a mediciones en una escala continua, de manera que no haya huecos o interrupciones.

Este capítulo analiza exclusivamente variables aleatorias discretas, pero los siguientes incluyen variables aleatorias continuas.



a) Variable aleatoria discreta: contador del número de asistentes a un cine.

b) Variable aleatoria continua: voltaje medido de una batería de un detector de humo.

FIGURA 4-2 Aparatos que se utilizan para contar y medir variables aleatorias discretas y continuas

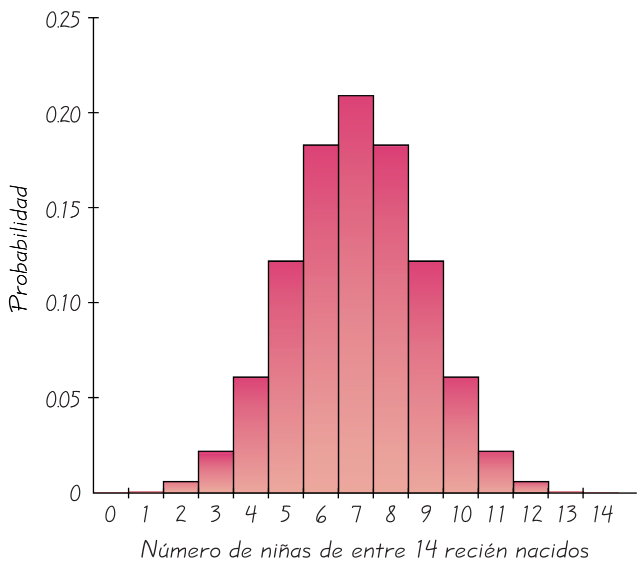
EJEMPLOS Los siguientes son ejemplos de variables aleatorias discretas y continuas:

1. Sea x = número de huevos que una gallina pone en un día. Ésta es una variable aleatoria *discreta*, porque sus únicos valores posibles son 0 o 1, o 2, etcétera. Ninguna gallina puede poner 2.343115 huevos, lo que sería posible si los datos provinieran de una escala continua.
2. El conteo del número de fanáticos que asiste a un concierto de, N Sync es un número entero y, por lo tanto, una variable aleatoria discreta. El aparato de conteo que se muestra en la figura 4-2a es capaz de indicar únicamente un número finito de valores, por lo que se utiliza para obtener valores de una variable aleatoria *discreta*.
3. Sea x = cantidad de leche que produce una vaca en un día. Ésta es una variable aleatoria *continua*, ya que puede tomar cualquier valor en un tramo continuo. En un solo día, una vaca llega a producir una cantidad de leche cuyo valor sería cualquiera entre 0 y 5 galones. Es posible obtener 4.123456 galones, debido a que la vaca no se restringe a las cantidades discretas de 0, 1, 2, 3, 4 o 5 galones.
4. La medida del voltaje de una batería de un detector de humo puede ser cualquier valor entre 0 y 9 voltios. Por lo tanto, se trata de una variable aleatoria continua. El voltímetro que se muestra en la figura 4-2b es capaz de indicar valores en una escala continua; por lo tanto, se utiliza para obtener valores de una variable aleatoria *continua*.

Gráficas

Hay varias formas para graficar una distribución de probabilidad, aquí consideraremos solamente al **histograma de probabilidad**. La figura 4-3 es un histograma de probabilidad muy similar al histograma de frecuencias relativas que se estudió

FIGURA 4-3 Histograma de probabilidad del número de niñas de entre 14 bebés recién nacidos



en el capítulo 2, pero la escala vertical muestra *probabilidades* en lugar de frecuencias relativas que se basan en resultados muestrales reales.

En la figura 4-3, observe que, a lo largo del eje horizontal, los valores de 0, 1, 2, . . . , 14 se localizan en el centro de los rectángulos, lo cual implica que cada uno de los rectángulos mide una unidad, de modo que las áreas de los rectángulos son 0.000, 0.001, 0.006, etcétera. Las áreas de estos rectángulos son iguales a las *probabilidades* en la tabla 4-1. En el capítulo 5 y en capítulos posteriores veremos que tal correspondencia entre el área y la probabilidad es muy útil en la estadística.

Toda distribución de probabilidad debe satisfacer cada uno de los dos requisitos siguientes:

Requisitos de una distribución de probabilidad

1. $\Sigma P(x) = 1$ donde x toma todos los valores posibles
2. $0 \leq P(x) \leq 1$ para cada valor individual de x

El primer requisito establece que la suma de las probabilidades de todos los valores posibles de la variable aleatoria debe ser igual a 1. Lo anterior tiene sentido cuando nos damos cuenta de que los valores de la variable aleatoria x representan todos los sucesos posibles en el espacio muestral completo, de modo que tenemos la certeza (con probabilidad 1) de que uno de los sucesos ocurrirá. En la tabla 4-1, la suma de todas las probabilidades es de 0.999; sería igual a 1 si elimináramos el pequeño error por redondeo ocupando más decimales. También, la regla de probabilidad que establece que $0 \leq P(x) \leq 1$ para cualquier suceso A (dado en la sección 3-2) implica que $P(x)$ debe estar entre 0 y 1 para cualquier valor de x. Nuevamente, remítase a la tabla 4-1 y observe que cada valor individual de $P(x)$ cae entre 0 y 1 para cualquier valor de x. Puesto que la tabla 4-1 satisface ambos requisitos, es un ejemplo de una distribución de probabilidad. Una distribución de probabilidad llega a describirse como una tabla, tal como la tabla 4-1; una gráfica, como la figura 4-3, o una fórmula.

EJEMPLO ¿Describirá la tabla 4-2 una distribución de probabilidad?

SOLUCIÓN Para ser una distribución de probabilidad, $P(x)$ debe satisfacer los dos requisitos anteriores. Pero

$$\begin{aligned}\Sigma P(x) &= P(0) + P(1) + P(2) + P(3)\\ &= 0.2 + 0.5 + 0.4 + 0.3\\ &= 1.4 \left[\text{que demuestra que } \Sigma P(x) \neq 1\right]\end{aligned}$$

Como no se satisface el primer requisito, concluimos que la tabla 4-2 no describe una distribución de probabilidad.

Tabla 4-2
Probabilidades de una variable aleatoria

x	$P(x)$
0	0.2
1	0.5
2	0.4
3	0.3

EJEMPLO ¿Determina si $P(x) = x/3$ (donde x puede ser 0, 1 o 2) es una distribución de probabilidad?

SOLUCIÓN Para la función dada, encontramos que $P(0) = 0/3$, $P(1) = 1/3$, y $P(2) = 2/3$, de manera que

1. $\Sigma P(x) = \frac{0}{3} + \frac{1}{3} + \frac{2}{3} = \frac{3}{3} = 1$
2. Cada uno de los valores $P(x)$ se encuentran entre 0 y 1.

Como ambos requisitos se satisfacen, la función $P(x)$ de este ejemplo es una distribución de probabilidad.

Media, varianza y desviación estándar

Recuerde que en el capítulo 2 describimos las siguientes características importantes de los datos (que pueden recordarse por medio de las siglas CVDDT "**C**uidado con los **V**irus que **D**estruyen **D**atos y **T**rabajo"):

1. *Centro:* valor representativo o promedio que indica la localización de la mitad del conjunto de los datos.
2. *Variación:* medida de la cantidad en que los valores de los datos varían entre sí.
3. *Distribución:* naturaleza o forma de la distribución de los datos (tales como normales, uniformes o sesgadas).
4. *Datos distantes:* valores muestrales que se alejan mucho de la vasta mayoría de los otros valores de la muestra.
5. *Tiempo:* características cambiantes de los datos a través del tiempo.

El histograma de probabilidad nos ofrece información acerca de la naturaleza o forma de la distribución. Además, podemos calcular la media, la varianza y la desviación estándar de los datos, los cuales proporcionan información acerca de otras características. La media, la varianza y la desviación estándar de una distribución de probabilidad se calculan aplicando las fórmulas 4-1, 4-2, 4-3 y 4-4.

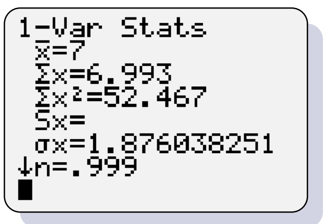
Fórmula 4-1 $\mu = \Sigma[x \cdot P(x)]$ media de una distribución de probabilidad

Fórmula 4-2 $\sigma^2 = \Sigma[(x - \mu)^2 \cdot P(x)]$ varianza de una distribución de probabilidad

Fórmula 4-3 $\sigma^2 = \Sigma[x^2 \cdot P(x)] - \mu^2$ varianza de una distribución de probabilidad

Fórmula 4-4 $\sigma = \sqrt{\Sigma[x^2 \cdot P(x)] - \mu^2}$ desviación estándar de una distribución de probabilidad

TI-83 Plus



Precaución: Evalúe $\Sigma[x^2 \cdot P(x)]$ elevando al cuadrado cada valor de x, multiplicando cada cuadrado por el $P(x)$ correspondiente y sumando.

La calculadora TI-83 Plus resulta útil para calcular la media y la desviación estándar. Aquí se muestra la representación visual de la calculadora TI-83 Plus, de la distribución de probabilidad descrita en la tabla 4-1. En la imagen de la TI-83 Plus, el valor que se muestra como $\bar{x}$ es en realidad el valor de la media μ, en tanto que el valor que se presenta como σx es el valor de la desviación estándar σ. Es decir, $\mu = 7$ y $\sigma = 1.876038251$.

Fundamentos de las fórmulas 4-1 a la 4-4

¿Por qué sirven las fórmulas 4-1 a la 4-4? La fórmula 4-1 logra lo mismo que la fórmula de la media para una tabla de frecuencias. (Recuerde que f representa la frecuencia de clase y N representa el tamaño de la población). Al rescribir la fórmula de la media de una tabla de frecuencias, de modo que se aplique a una población, y luego cambiando su forma, obtendremos

$$\mu = \frac{\Sigma(f \cdot x)}{N} = \Sigma \frac{f \cdot x}{N} = \Sigma x \cdot \frac{f}{N} = \Sigma x \cdot P(x)$$

En la fracción f/N, el valor de f es la frecuencia con que ocurre el valor x y N es el tamaño de la población, de modo que f/N es la probabilidad del valor de x.

Un razonamiento similar nos permite considerar la fórmula de la varianza del capítulo 2 y aplicarla a una variable aleatoria para una distribución de probabilidad; el resultado es la fórmula 4-2. La fórmula 4-3 es una versión abreviada que siempre producirá el mismo resultado que la fórmula 4-2. Aunque la fórmula 4-3 es más fácil de usar, la fórmula 4-2 es más sencilla de comprender directamente. Con base en la fórmula 4-2, expresamos la desviación estándar como

$$\sigma = \sqrt{\Sigma(x - \mu)^2 \cdot P(x)}$$

o como la forma equivalente dada en la fórmula 4-4.

Cuando utilice las fórmulas 4-1 a la 4-4, aplique esta regla para redondear los resultados.

Regla de redondeo para μ, σ y σ^2

Redondee los resultados llevando una posición decimal más que el número de posiciones decimales utilizadas para la variable aleatoria x. Si los valores de x son enteros, redondee μ, σ y σ^2 a una posición decimal.

En ocasiones, es necesario usar una regla diferente de redondeo debido a circunstancias especiales, tales como resultados que requieren más decimales para tener un significado. Por ejemplo, para aviones de propulsión a chorro de cuatro motores, el número medio de motores que funcionan exitosamente durante un vuelo es de 3.999714286, que se convierte en 4.0 cuando se redondea a una posición decimal más que los datos originales. Aquí, el 4.0 sería confuso, ya que sugiere que todos los motores de aviones de propulsión a chorro siempre funcionan bien. Necesitamos más precisión para reflejar correctamente la media verdadera, como la precisión en el número 3.999714.

Identificación de resultados poco comunes con la regla práctica del intervalo

La regla práctica del intervalo (que se estudió en la sección 2-5), también resulta útil para interpretar los valores de una desviación estándar. Según la regla práctica del intervalo, la mayoría de los valores deben caer dentro de dos desviaciones estándar de la media; no es frecuente que un valor difiera de la media en más de dos desviaciones estándar. (Por lo general, el uso de dos desviaciones estándar no es un valor absolutamente rígido; en su lugar, se utilizan otros valores, como el 3). De esta manera, identificamos valores “poco comunes” si se determinara que caen fuera de tales límites:

$$\text{valor máximo común} = \mu + 2\sigma$$
$$\text{valor mínimo común} = \mu - 2\sigma$$

EJEMPLO La tabla 4-1 describe la distribución de probabilidad del número de niñas entre 14 bebés recién nacidos que se seleccionaron aleatoriamente. Suponiendo que repetimos el estudio, seleccionamos aleatoriamente 14 bebés recién nacidos y contamos el número de niñas en cada ocasión, calcule el número medio de niñas (de entre 14), la varianza y la desviación estándar. Utilice dichos resultados y la regla práctica del intervalo para obtener los valores máximo y mínimo comunes.

SOLUCIÓN En la tabla 4-3, las dos columnas a la izquierda describen la distribución de probabilidad que se presentó en la tabla 4-1; elaboramos las tres columnas a la derecha con el propósito de lograr los cálculos requeridos.

Tabla 4-3 Cálculo de μ, σ y σ^2 para una distribución de probabilidad

x	$P(x)$	$x \cdot P(x)$	x^2	$x^2 \cdot P(x)$
0	0.000	0.000	0	0.000
1	0.001	0.001	1	0.001
2	0.006	0.012	4	0.024
3	0.022	0.066	9	0.198
4	0.061	0.244	16	0.976
5	0.122	0.610	25	3.050
6	0.183	1.098	36	6.588
7	0.209	1.463	49	10.241
8	0.183	1.464	64	11.712
9	0.122	1.098	81	9.882
10	0.061	0.610	100	6.100
11	0.022	0.242	121	2.662
12	0.006	0.072	144	0.864
13	0.001	0.013	169	0.169
14	0.000	0.000	196	0.000
Total		6.993 ↑ $\Sigma[x \cdot P(x)]$		52.467 ↑ $\Sigma[x^2 \cdot P(x)]$

¿Es seguro el paracaidismo?

De las más de 100,000 personas que realizan cerca de 2.25 millones de saltos en paracaídas, aproximadamente 30 mueren cada año. En comparación, un año típico incluye alrededor de 200 muertes en el buceo, 7,000 ahogamientos, 900 muertes en bicicletas, 800 muertes por relámpagos y 1,150 muertes por picaduras de abeja. Desde luego, tales cifras no significan necesariamente que el paracaidismo sea más seguro que andar en bicicleta o la natación. En una comparación justa, deben incluirse tasas de mortalidad, no sólo el número total de fallecimientos.

El autor, con gran osadía, realizó dos saltos en paracaídas, pero desistió después de no caer dentro de la amplia zona de aterrizaje, en ambas ocasiones. También voló en una ala delta, un globo aerostático y en el dirigible Goodyear.

Con el uso de las fórmulas 4-1 y 4-3, así como con los resultados de la tabla, obtendremos

$$\mu = \Sigma[x \cdot P(x)] = 6.993 = 7.0 \qquad \text{(redondeado)}$$
$$\sigma^2 = \Sigma[x^2 \cdot P(x)] - \mu^2$$
$$= 52.467 - 6.993^2 = 3.564951 = 3.6 \qquad \text{(redondeado)}$$

La desviación estándar es la raíz cuadrada de la varianza, entonces

$$\sigma = \sqrt{3.564951} = 1.9 \qquad \text{(redondeado)}$$

Sabemos que entre los grupos de 14 bebés recién nacidos, el número medio de niñas es de 7.0, la varianza es de 3.6 "niñas cuadradas" y la desviación estándar es de 1.9 niñas.

Con la aplicación de la regla práctica del intervalo, ahora calculamos los valores máximo y mínimo comunes de la siguiente manera:

$$\text{valor máximo común: } \mu + 2\sigma = 7.0 + 2(1.9) = 10.8$$
$$\text{valor mínimo común: } \mu - 2\sigma = 7.0 - 2(1.9) = 3.2$$

INTERPRETACIÓN Con base en estos resultados, concluimos que para grupos de 14 bebés que se seleccionaron aleatoriamente, el número de niñas debe caer comúnmente entre 3.2 y 10.8.

Identificación de resultados poco comunes con probabilidades

Recomendación importante: La siguiente explicación incluye algunos conceptos difíciles, pero también un método sumamente importante, que se utiliza con frecuencia en la estadística. Debe hacer su mejor esfuerzo para comprender dicha explicación y leerla varias veces si es necesario. Tenga en mente que tal explicación se basa en la *regla del suceso poco común* que se estudió en la sección 3-2:

Si, bajo un supuesto dado (como el de que niñas y niños son igualmente probables), la probabilidad de un suceso particular que se observa (como 13 niñas en 14 nacimientos) es extremadamente pequeña, concluimos que el supuesto probablemente no sea correcto.

En el problema del capítulo señalamos que, con la técnica MicroSort, resultaron 13 niñas entre 14 bebés. ¿Será el resultado poco común? ¿Sugiere que en realidad la técnica es efectiva, o podrían resultar 13 niñas entre 14 bebés sólo por azar? Para resolver esto, utilizamos la regla práctica del intervalo con la finalidad de calcular los resultados probables máximo y mínimo. Pero aquí consideraremos otro método: calcularemos la probabilidad de tener 13 *o más* niñas (no la de tener *exactamente* 13 niñas). Es difícil entender por qué la probabilidad de 13 *o más niñas* es la probabilidad relevante; por lo tanto, trataremos de aclararlo con un ejemplo más sencillo.

Suponga que lanza una moneda para determinar si se ven favorecidas las caras y que 1000 lanzamientos dan como resultado 501 caras, lo cual no es evidencia de que la moneda favorezca las caras, ya que es muy fácil obtener un resultado de 501 caras en 1000 lanzamientos al azar. Sin embargo, la probabilidad de obtener *exactamente* 500 caras en 1000 lanzamientos es bastante baja: 0.0252. Esa baja probabilidad refleja el hecho de que, con 1000 lanzamientos, cada número *es-*

Elección de números de la lotería

En una lotería estatal tradicional, usted selecciona seis números diferentes. Después de una selección aleatoria, los boletos con la combinación correcta comparten el premio. Puesto que los números ganadores se seleccionan aleatoriamente, cualquier elección de seis números tendrá la misma posibilidad, pero algunas combinaciones son mejores que otras. La combinación de 1, 2, 3, 4, 5, 6 es una mala elección, ya que muchas personas tienden a seleccionarla. En una lotería de Florida, con un premio de 105 millones de dólares, 52,000 boletos incluían 1, 2, 3, 4, 5, 6; si tal combinación hubiera ganado, el premio hubiese sido de tan sólo 1,000 dólares. Es más sensato elegir combinaciones que muchas otras personas no seleccionan. Evite combinaciones que forman un patrón en el boleto.

pecífico de caras tendrá una probabilidad sumamente baja. Sin embargo, no consideramos que 500 caras en 1000 lanzamientos sea *poco común*, puesto que la probabilidad de obtener *al menos* 501 caras es alta: 0.487. Este principio se generaliza de la siguiente manera:

Uso de las probabilidades para determinar resultados poco comunes

- **Inusualmente alto:** x éxitos en n ensayos es un número *inusualmente alto* de éxitos, si $P(x$ o más) es muy pequeño (como 0.05 o menos).
- **Inusualmente bajo:** x éxitos en n ensayos es un número *inusualmente bajo* de éxitos, si $P(x$ o menos) es muy pequeño (como 0.05 o menos).

PROBLEMA 4 DEL CAPÍTULO

EJEMPLO Selección del género Con el uso de los dos criterios anteriores que se basan en probabilidades, ¿será poco común que resulten 13 niñas en 14 nacimientos? ¿Parecería que la técnica de MicroSort de la selección del género es efectiva?

SOLUCIÓN Trece niñas entre 14 nacimientos es inusualmente alto si P (13 o más niñas) es muy pequeño. Si nos remitimos a la tabla A-1, obtendremos el siguiente resultado:

$$\begin{aligned} P\,(13 \text{ o más niñas}) &= P(13) + P(14) \\ &= 0.001 + 0.000 \\ &= 0.001 \end{aligned}$$

INTERPRETACIÓN Puesto que la probabilidad de 0.001 es demasiado baja, concluimos que es poco común que resulten 13 niñas en 14 nacimientos. Lo anterior sugiere que la técnica MicroSort de selección del género parece ser efectiva, ya que es altamente improbable que el resultado de 13 niñas en 14 nacimientos suceda por azar.

Valor esperado

La media de una variable aleatoria discreta es el resultado medio teórico de un número infinito de ensayos. Podemos considerar esa media como el *valor esperado* en el sentido de que constituye el valor promedio que esperaríamos obtener si los ensayos pudiesen continuar de manera indefinida. Los usos del valor esperado (también se le llama *esperanza* o *esperanza matemática*), que son extensos y variados, juegan un papel muy importante en una área de aplicación que se denomina *teoría de decisión*.

Definición

Valor esperado (de una variable aleatoria discreta): se denota con E y representa el valor promedio de los resultados. Se obtiene calculando el valor de $\Sigma[x \cdot P(x)]$.

$$E = \Sigma[x \cdot P(x)]$$

Con la fórmula 4-1 vemos que $E = \mu$; es decir, la media de una variable aleatoria discreta es la misma que su valor esperado. Repita el procedimiento de lanzar una moneda cinco veces; el número *medio* de caras es de 2.5; cuando se lanza una moneda cinco veces, el *valor esperado* del número de caras es también de 2.5.

EJEMPLO El juego Pick 3 de Nueva Jersey Hace años, miembros de grupos del crimen organizado realizaban juegos de números. En la actualidad, muchos miembros de gobiernos organizados, así como algunos gobiernos no tan bien organizados, realizan juegos similares. El juego Pick 3 de Nueva Jersey es muy similar a los antiguos juegos ilegales. Una apuesta "legal" funciona de la siguiente manera: apueste 50 centavos y seleccione un número de tres dígitos entre 000 y 999. Si sus tres dígitos coinciden con los que se seleccionan aleatoriamente, recibiría 275 dólares, así que obtendría una ganancia neta de 274.50 dólares (puesto que no se le devuelven sus 50 centavos de apuesta). Suponga que apuesta 50 centavos al número 007, ¿cuál es el valor esperado de ganar o perder?

SOLUCIÓN Para esta apuesta hay dos resultados sencillos: gana o pierde. Como usted eligió el número 007 y hay 1000 posibilidades (desde 000 hasta 999), su probabilidad de ganar es de $1/1000$ (o 0.001) y la de perder es de $999/1000$ (o 0.999). La tabla 4-4 resume dicha situación.

En la tabla 4-4 observamos que cuando apostamos 50 centavos en el juego Pick 3 de Nueva Jersey, nuestro valor esperado es

$$E = \Sigma[x \cdot P(x)] = -22.5 \text{ centavos}$$

El resultado puede interpretarse de la siguiente manera: a la larga, por cada apuesta de 50 centavos podemos esperar perder un promedio de 22.5 centavos. En cualquier juego individual perderá 50 centavos o ganará 274.50 dólares. Aunque no es posible que usted pierda 22.5 centavos en un juego individual, el valor esperado de −22.5 centavos muestra que, después de muchos juegos, la pérdida promedio por juego es de 22.5 centavos.

Tabla 4-4 Juego Pick 3 de Nueva Jersey

Suceso	x	$P(x)$	$x \cdot P(x)$
Ganancia	\$274.50	0.001	\$0.2745
Pérdida	−\$0.50	0.999	−\$0.4995
Total			−\$0.225 (o −22.5¢)

En esta sección aprendimos que una variable aleatoria tiene un valor numérico asociado a cada resultado de algún procedimiento aleatorio, así como que una distribución de probabilidad tiene una probabilidad asociada a cada valor de una variable aleatoria. Examinamos métodos para calcular la media, la varianza y la desviación estándar de una distribución de probabilidad. Vimos que el valor esperado de una variable aleatoria es, en realidad, igual a la media. También aprendimos que no debemos esperar enriquecernos con el juego de lotería Pick 3 de Nueva Jersey.

4-2 Destrezas y conceptos básicos

Identificación de variables aleatorias discretas continuas. *En los ejercicios 1 y 2, identifique si la variable aleatoria dada es discreta o continua.*

1. **a.** La estatura de un jugador de basquetbol de la NBA, que se selecciona aleatoriamente.
b. El número de puntos que anota en una temporada un jugador de basquetbol de la NBA, seleccionado aleatoriamente.
c. El tiempo exacto de juego de un jugador de basquetbol de la NBA, que se selecciona aleatoriamente.
d. El número de atletas que participaron en cualquier juego de la NBA en una temporada.
e. El salario de un jugador de basquetbol de la NBA, seleccionado aleatoriamente.

2. **a.** El costo de la realización de una película que se selecciona aleatoriamente.
b. El número de películas que actualmente se exhiben en los cines de Estados Unidos.
c. La duración exacta de una película seleccionada aleatoriamente.
d. El número de actores que aparecen en una película que se selecciona aleatoriamente.
e. El peso del actor principal de una película seleccionada aleatoriamente.

Identificación de distribuciones de probabilidad. *En los ejercicios 3 a 10, determine si se trata de una distribución de probabilidad. En los casos en que no se describa una distribución de probabilidad, identifique los requisitos que no se satisfacen. En los casos en que se describa una distribución de probabilidad, calcule su media y desviación estándar.*

3. Selección de género En un estudio con el método MicroSort de selección del género, las parejas de un grupo control no reciben tratamiento y cada una de ellas tiene tres hijos. La distribución de probabilidad del número de niñas se presenta en la tabla anexa.

x	$P(x)$
0	0.125
1	0.375
2	0.375
3	0.125

4. Control de calidad de DVD Durante la fabricación del DVD de Sony, se seleccionan aleatoriamente grupos de DVD y se calcula el número de defectos x en cada grupo.

x	$P(x)$
0	0.502
1	0.365
2	0.098
3	0.011
4	0.001

5. Renta de videocintas La tabla adjunta se construye de datos que se obtienen en un estudio del número de videocintas rentadas en Blockbuster.

x	$P(x)$
0	0.04
1	0.26
2	0.36
3	0.20
4	0.08

6. Seguro de vida La compañía Telektronic brinda pólizas de seguro de vida a sus cuatro ejecutivos principales; la variable aleatoria x es el número de estos empleados que sobreviven durante el año siguiente.

x	$P(x)$
0	0.0000
1	0.0001
2	0.0006
3	0.0387
4	0.9606

7. Sentencias previas Se selecciona aleatoriamente a un preso, convicto por manejar intoxicado; la distribución de probabilidad del número x de sentencias previas por este delito se describe en la tabla adjunta (según datos del Departamento de Justicia de Estados Unidos).

x	$P(x)$
0	0.512
1	0.301
2	0.132
3	0.055

8. Vuelos sobresaturados La línea Air America tiene, por rutina, la política de vender un número de boletos que rebasa el número de asientos por cada vuelo, ya que la experiencia demuestra que algunos pasajeros no se presentan. La variable aleatoria x representa el número de pasajeros que no pueden abordar porque hay más pasajeros que asientos.

x	$P(x)$
0	0.805
1	0.113
2	0.057
3	0.009
4	0.002

9. Número de juegos en una Serie Mundial de Beisbol Con base en resultados que se encontraron en el *Information Please Almanac*, hay una probabilidad del 0.1809 de que la Serie Mundial de Beisbol dure cuatro juegos, una probabilidad del 0.2234 de que dure cinco juegos, una probabilidad de 0.2234 de que dure seis juegos y una probabilidad del 0.3723 de que dure siete juegos. ¿Será poco común que un equipo se corone ganando cuatro juegos seguidos?

10. Reconocimiento de marca En un estudio de reconocimiento de la marca Sony, se entrevistaron grupos de cuatro consumidores. Si x es el número de personas en el grupo que reconocen la marca Sony, entonces x puede ser 0, 1, 2, 3 o 4, en tanto que las probabilidades correspondientes son 0.0016, 0.0250, 0.1432, 0.3892 y 0.4096. ¿Será poco común seleccionar aleatoriamente cuatro consumidores y descubrir que ninguno de ellos reconoce la marca Sony?

11. Cálculo del valor esperado en los dados Cuando usted apuesta 5 dólares en un casino a “pasar la línea” en el juego de dados, hay una probabilidad de $244/495$ de que gane \$5 y una probabilidad de $251/495$ de que pierda \$5. ¿Cuál es su valor esperado? A la larga, ¿cuánto pierde por cada dólar que apueste?

12. Cálculo del valor esperado en la ruleta Cuando usted apuesta \$5 en un casino al número 7 en la ruleta, tiene una probabilidad de $1/38$ de ganar \$175, y una probabilidad de $37/38$ de perder \$5. Si apuesta \$5 a que el resultado es un número impar, la probabilidad de ganar \$5 es de $18/38$ y la probabilidad de perder \$5 es de $20/38$.
   a. Si apuesta \$5 al número 7, ¿cuál es su valor esperado?
   b. Si apuesta \$5 a que el resultado es un número impar, ¿cuál es su valor esperado?
   c. ¿Cuál de estas opciones es mejor: apostar al 7, apostar a número impar o no apostar? ¿Por qué?

13. Cálculo del valor esperado para una póliza de seguro de vida La compañía de seguros CNA le cobra a Mike \$250 por un año de una póliza de seguro de vida de \$100,000. Puesto que Mike es un hombre de 21 años, hay una probabilidad del 0.9985 de que sobreviva durante un año (según datos del National Center for Health Statistics de Estados Unidos).
   a. Desde la perspectiva de Mike, ¿cuáles son los valores de los dos resultados diferentes?
   b. Si Mike compra la póliza, ¿cuál es su valor esperado?
   c. ¿Cuál sería el costo de la póliza del seguro si la compañía sale a mano (a la larga eso sucede con muchas pólizas), en lugar de obtener una ganancia?
   d. Puesto que el valor esperado de Mike es negativo (de modo que la compañía obtiene una ganancia), ¿por qué es aconsejable que Mike o cualquiera otra persona adquiera un seguro de vida?

14. Cálculo del valor esperado de la rifa organizada por una revista Recientemente, la revista *Reader’s Digest* realizó una rifa en la que los premios se listaron junto con las

probabilidades de ganar: $1,000,000 (1 posibilidad en 90,000,000); $100,000 (1 posibilidad en 110,000,000); $25,000 (1 posibilidad en 110,000,000); $5000 (1 posibilidad en 36,667,000), y $2500 (1 posibilidad en 27,500,000).

a. Calcule el valor esperado de la cantidad a ganar con un boleto.
b. Calcule el valor esperado si el costo de un boleto en esta rifa equivale al de una estampilla postal. ¿Vale la pena participar en esta rifa?

15. Cálculo del valor esperado del juego Pick 4 de Nueva Jersey En el juego de lotería Pick 4 de Nueva Jersey, usted paga 5 centavos para seleccionar una secuencia de cuatro dígitos, como 7273. Si usted gana al seleccionar la misma secuencia de los cuatro dígitos resultantes, obtiene $2788.

a. ¿Cuántas selecciones distintas son posibles?
b. ¿Cuál es la probabilidad de ganar?
c. Si gana, ¿cuál es su ganancia neta?
d. Calcule el valor esperado.
e. Uno de los ejemplos de esta sección demostró que si apuesta 50 centavos en el juego Pick 3 de Nueva Jersey, su valor esperado es -22.5 centavos. ¿Cuál juego es mejor: Pick 3 o Pick 4 de Nueva Jersey? ¿Por qué?

16. Cálculo del valor esperado del juego Pick 3 de Illinois En el juego Pick 3 de Illinois, usted paga 50 centavos para seleccionar una secuencia de tres dígitos, como 911. Si usted gana al seleccionar la misma secuencia de los tres dígitos resultantes, obtendría $250.

a. ¿Cuántas selecciones distintas son posibles?
b. ¿Cuál es la probabilidad de ganar?
c. Si gana, ¿cuál es su ganancia neta?
d. Calcule el valor esperado.
e. Uno de los ejemplos de esta sección demostró que si apuesta 50 centavos en el juego Pick 3 de Nueva Jersey, su valor esperado es -22.5 centavos. ¿Cuál juego es mejor: Pick 3 de Illinois o Pick 3 de Nueva Jersey? ¿Por qué?

17. Determinación de la eficacia de la técnica de selección del género Suponga que, en una prueba de una técnica de selección del género, un ensayo clínico da como resultado nueve niñas en 14 nacimientos. Remítase a la tabla 4-1 y obtenga las probabilidades que se indican.

a. Calcule la probabilidad de que resulten exactamente nueve niñas en 14 nacimientos.
b. Calcule la probabilidad de tener nueve o más niñas en 14 nacimientos.
c. ¿Qué probabilidad es relevante para determinar si nueve niñas en 14 nacimientos es un suceso inusualmente alto: el resultado del inciso *a* o el del inciso *b*?
d. ¿Sugerirá el resultado de nueve niñas en 14 nacimientos que la técnica de selección del género es efectiva? ¿Por qué?

18. Determinación de la eficacia de la técnica de selección del género Suponga que, en una prueba de una técnica de selección del género, un ensayo clínico da como resultado 12 niñas en 14 nacimientos. Remítase a la tabla 4-1 y obtenga las probabilidades que se indican.

a. Calcule la probabilidad de tener exactamente 12 niñas en 14 nacimientos.
b. Calcule la probabilidad de tener 12 o más niñas en 14 nacimientos.
c. ¿Qué probabilidad es relevante para determinar si 12 niñas en 14 nacimientos es un suceso inusualmente alto: el resultado del inciso *a* o del inciso *b*?
d. ¿Sugerirá el resultado de 12 niñas en 14 nacimientos que la técnica de selección del género es efectiva? ¿Por qué?

19. Determinación de la eficacia de la técnica de selección del género Suponga que, en una prueba de una técnica de selección del género, un ensayo clínico tiene como resultado 11 niñas en 14 nacimientos. Remítase a la tabla 4-1 y obtenga las probabilidades que se indican.

continúa

a. ¿Cuál es el valor de probabilidad que debe utilizarse para determinar si el resultado de 11 niñas en 14 nacimientos es inusualmente alto?
b. ¿Sugerirá el resultado de 11 niñas en 14 nacimientos que la técnica de selección del género es efectiva? ¿Por qué?

20. Determinación de la eficacia de la técnica de selección del género Suponga que, en una prueba de una técnica de selección del género, un ensayo clínico tiene como resultado 10 niñas en 14 nacimientos. Remítase a la tabla 4-1 y obtenga las probabilidades que se indican.
a. ¿Cuál es el valor de probabilidad que debe utilizarse para determinar si el resultado de 10 niñas en 14 nacimientos es inusualmente alto?
b. ¿Sugerirá el resultado de 10 niñas en 14 nacimientos que la técnica de selección del género es efectiva? ¿Por qué?

21. Poderes psíquicos Bob, quien se considera un "mentalista", afirma que puede pasar una prueba de verdadero/falso adivinando. Para comprobar su afirmación, se le plantean 14 preguntas con respuestas de verdadero/falso, de las cuales, ocho responde correctamente. Bob asevera que obtener ocho respuestas correctas, en 14 preguntas, es una prueba de sus poderes especiales, ya que obtuvo más respuestas correctas que las siete que se esperarían por el azar. ¿Es válida la aseveración de Bob? ¿Por qué? (*Sugerencia:* Para calcular la probabilidad que se requiere, remítase a la tabla 4-1 y cambie "niñas" por "correcto". Las probabilidades de niñas serán las mismas probabilidades de conjeturas correctas).

22. Poderes psíquicos Bob, quien se considera un "mentalista", afirma que es capaz de pasar una prueba de verdadero/falso adivinando. Para comprobar su afirmación, se le plantean 14 preguntas con respuestas de verdadero/falso, de las cuales dos responde correctamente. ¿Será poco común el número de respuestas correctas? ¿Por qué? ¿Será válida la afirmación de Bob? (*Sugerencia*: Para calcular la probabilidad que se requiere, remítase a la tabla 4-1 y cambie "niñas" por "correcto". Las probabilidades de niñas serán las mismas probabilidades de conjeturas correctas).

4-2 Más allá de lo básico

23. Bonos especulativos Kim Hunter tiene $1000 para invertir, por lo que su analista financiero le recomienda dos tipos de bonos especulativos. Los bonos A tienen un rendimiento anual del 6%, con una tasa de incumplimiento del 1%. Los bonos B tienen un rendimiento anual del 8%, con una tasa de incumplimiento del 5%. (Si el bono incumple, se pierden los $1000). ¿Cuál de los dos bonos es mejor? ¿Por qué? ¿Debe ella elegir cualquiera de los dos bonos? ¿Por qué?

24. Cálculo de la media y la desviación estándar Sea x la variable aleatoria, la cual representa el número de niñas en una familia de cuatro hijos. Construya una tabla que describa la distribución de probabilidad; después, calcule la media y la desviación estándar. *(Sugerencia*: Liste los distintos resultados posibles).

25. Partes defectuosas: cálculo de la media y la desviación estándar Sky Ranch es un proveedor de partes para aeronaves. Sus existencias incluyen ocho altímetros que están correctamente calibrados y dos que no lo están. Se seleccionan tres altímetros aleatoriamente sin reemplazo. Sea x la variable aleatoria, la cual representa el número de aparatos que no están calibrados correctamente. Calcule la media y la desviación estándar de la variable aleatoria x.

26. Números generados por computadora, transformados a puntuación z Con frecuencia se utilizan computadoras para generar aleatoriamente los últimos dígitos de números telefónicos de sujetos potenciales a encuestar. Los dígitos se seleccionan de manera que todos sean igualmente probables. La variable aleatoria x es el número elegido.

continúa

a. Calcule la media en la desviación estándar de la variable aleatoria x.
b. Calcule la puntuación z para cada valor de x; después, la media y la desviación estándar de las puntuaciones z.
c. ¿Resultarán la misma media y la desviación estándar del inciso b de cada distribución de probabilidad?

27. Enteros con la misma probabilidad: media y desviación estándar Suponga que una distribución de probabilidad está descrita por la variable aleatoria discreta x, que puede tomar los valores 1, 2, . . . , n, y que dichos valores son igualmente probables. La distribución de probabilidad tiene una media y una desviación estándar como se describe a continuación:

$$\mu = \frac{n+1}{2} \quad \text{y} \quad \sigma = \sqrt{\frac{n^2-1}{12}}$$

a. Demuestre que $\mu = (n+1)/2$ para el caso de $n = 5$.
b. Demuestre que $\sigma = \sqrt{(n^2-1)/12}$ para el caso de $n = 5$.
c. Con la finalidad de probar a un individuo que afirma tener poderes extrasensoriales, usted selecciona aleatoriamente números enteros entre 1 y 20, en tanto que la variable aleatoria x es el número que se selecciona. Calcule la media y la desviación estándar de x.

28. Dados con marcas que permitan una distribución uniforme Suponga que tiene dos dados en blanco, de modo que puede marcar las 12 caras con los números que desee. Describa de qué manera marcaría los dados para que, cuando tire ambos, los totales de los dos dados se distribuyan de manera uniforme y cada uno de los resultados de 1, 2, 3, . . . , 12 tenga una probabilidad de 1/12. (Véase "Can One Load a Set of Dice so that the Sum is Uniformly Distributed?", de Chen, Rao y Shreve, *Mathematics Magazine,* vol. 70, núm. 3).

4-3 Distribuciones de probabilidad binomial

En la sección 4-2 estudiamos diversas distribuciones discretas de probabilidad, pero en esta sección nos enfocaremos en un tipo específico: la distribución de probabilidad binomial. Las distribuciones de probabilidad binomial son importantes porque nos permiten enfrentar circunstancias en las que los resultados pertenecen a *dos* categorías relevantes, tales como productos aceptables/defectuosos o respuestas sí/no a una pregunta de encuesta. El problema del capítulo implica el conteo del número de niñas en 14 nacimientos; incluye las dos categorías niño/niña, por lo que posee el elemento fundamental requerido de "dualidad". En la siguiente definición se plantean otros requisitos.

Definición

Distribución de probabilidad binomial: resulta de un procedimiento que cumple con todos los requisitos siguientes:

1. El procedimiento tiene un *número fijo de ensayos*.
2. Los ensayos deben ser *independientes*. (El resultado de cualquier ensayo individual no afecta las probabilidades de los otros ensayos).
3. Todos los resultados de cada ensayo deben estar clasificados en *dos categorías*.
4. Las probabilidades tienen que mantenerse *constantes* para cada ensayo.

Si un procedimiento satisface los cuatro requisitos, la distribución de la variable aleatoria x se denomina *distribución de probabilidad binomial* (o *distribución binomial*). Suele usarse la notación siguiente:

Notación para distribuciones de probabilidad binomial

E y F (éxito y fracaso) denotan las dos categorías posibles de todos los resultados; p y q denotan las probabilidades de E y F, respectivamente, de modo que

$$P(\text{E}) = p \qquad (p = \text{probabilidad de un éxito})$$

$$P(\text{F}) = 1 - p = q \qquad (q = \text{probabilidad de un fracaso})$$

n	denota el número fijo de ensayos.
x	denota un número específico de éxitos en n ensayos, de modo que x puede ser cualquier número entero entre 0 y n, inclusive.
p	denota la probabilidad de *éxito* en *uno* de n ensayos.
q	denota la probabilidad de *fracaso* en *uno* de n ensayos.
$P(x)$	denota la probabilidad de lograr exactamente x éxitos en los n ensayos.

El término *éxito* que se utiliza aquí es arbitrario y no necesariamente representa algo bueno. Cualquiera de las dos categorías posibles puede denominarse el éxito E, siempre y cuando su probabilidad se identifique como p. Una vez que se designa una categoría como éxito E, asegúrese de que p es la probabilidad de un éxito y que x es el número de éxitos. Es decir, asegúrese de que los valores p y x se refieran a la misma categoría que se designa como un éxito. (El valor de q se puede calcular siempre al restar p de 1; si $p = 0.95$, entonces $q = 1 - 0.95 = 0.05$). He aquí una sugerencia importante para trabajar con problemas de probabilidad binomial:

Asegúrese de que x y p se refieran a la *misma* categoría que se denomina como un éxito.

Cuando seleccionamos una muestra para algún análisis estadístico, por lo general realizamos el muestreo sin reemplazo. Por ejemplo, al probar artículos manufacturados o realizar encuestas, solemos diseñar el proceso de muestreo de modo que los artículos elegidos no se seleccionan una segunda vez. Estrictamente hablando, el muestreo sin reemplazo implica sucesos dependientes, que violan el segundo requisito de la definición anterior. Sin embargo, la siguiente regla práctica se basa en el hecho de que si la muestra es muy pequeña, en relación con el tamaño de la población, podemos tratar a los ensayos como independientes (aun cuando en realidad sean dependientes), ya que la diferencia en los resultados será insignificante.

Cuando se realiza un muestreo sin reemplazo, los sucesos pueden tratarse como si fueran independientes, si el tamaño de la muestra no es mayor que el 5% del tamaño de la población. (Es decir, $n \leq 0.05N$).

Profetas de las ganancias

Muchos libros y programas de computadoras aseguran ser útiles para predecir números ganadores de la lotería. Algunos utilizan la teoría de que ciertos números están "rezagados" (y hay que seleccionarlos), ya que no han salido con frecuencia; otros creen en la teoría de que algunos números son "fríos" (y deben evitarse), puesto que no han salido con frecuencia; incluso existen más que utilizan la astrología, la numerología o los sueños. Ya que la selección de las combinaciones de números ganadores de la lotería son sucesos independientes, dichas teorías son inútiles. Un método válido es el de elegir números que sean "raros", en el sentido de que no son seleccionados por otras personas, de forma que si usted gana, no se le obligará a compartir sus ganancias con otros. Por tal razón, la combinación de 1, 2, 3, 4, 5 y 6 es inadecuada, puesto que muchos individuos la utilizan, mientras que la combinación 12, 17, 18, 33, 40, 46 es mucho mejor, al menos hasta la publicación de este libro.

EJEMPLO Análisis de respuestas de opción múltiple Por su facilidad de corrección, las preguntas de opción múltiple se utilizan con frecuencia para realizar exámenes en el salón de clases, para la prueba SAT, para la prueba MCAT en las escuelas de medicina, la prueba LSAT en las escuelas de leyes y en muchas otras circunstancias. Un profesor que imparte un curso de psicología del comportamiento anormal planea aplicar un examen sorpresa que consta de cuatro preguntas de opción múltiple, cada una con cinco respuestas posibles (a, b, c, d, e), pero sólo una de ellas correcta. Supongamos que un estudiante sin preparación adecuada hace adivinanzas al azar y que queremos calcular la probabilidad de que tenga exactamente tres respuestas correctas en las cuatro preguntas.

a. ¿Resultará este proceso en una distribución binomial?

b. Si el proceso resulta en una distribución binomial, identifique los valores de n, x, p y q.

SOLUCIÓN

a. El procedimiento sí satisface los requisitos de una distribución binomial, como se muestra a continuación:

1. El número de ensayos (4) es fijo.
2. Los cuatro ensayos son independientes, ya que una respuesta correcta o incorrecta para cada pregunta individual no afecta la probabilidad de que otra pregunta sea correcta o incorrecta.
3. Cada uno de los cuatro ensayos tiene dos categorías de resultados posibles: correcta o incorrecta.
4. Para cada pregunta hay cinco respuestas posibles (a, b, c, d, e), pero sólo una de ellas es correcta, por lo que la probabilidad de una respuesta correcta es de $1/5$ o 0.2. La probabilidad de cada uno de los cuatro ensayos permanece constante.

b. Habiendo concluido que el procedimiento dado sí resulta en una distribución binomial, procedamos a identificar los valores de n, x, p y q.

1. Con cuatro preguntas de un examen, tenemos que $n = 4$.
2. Buscamos la probabilidad de exactamente tres respuestas correctas; entonces, $x = 3$.
3. La probabilidad de éxito (respuesta correcta) para una pregunta es de 0.2; por lo tanto, $p = 0.2$.
4. La probabilidad de fracaso (respuesta incorrecta) es de 0.8; por lo tanto, $q = 0.8$.

Nuevamente, es muy importante asegurarse de que tanto x como p se refieren al mismo concepto de "éxito". En este ejemplo, usamos x para contar las respuestas *correctas*, de modo que p debe ser la probabilidad de una respuesta *correcta*. Por consiguiente, x y p sí utilizan aquí el mismo concepto de éxito (respuesta correcta).

Ahora presentaremos tres métodos para calcular las probabilidades correspondientes a la variable aleatoria x en una distribución binomial. El primero, que implica cálculos empleando la *fórmula de probabilidad binomial*, es la base de los otros dos. El segundo implica el uso de la tabla A-1; y el tercero, el uso de un programa de cómputo o una calculadora. Si está utilizando alguna de estas dos herramientas que producen, de forma automática, probabilidades binomiales, le recomendamos que resuelva uno o dos ejercicios por medio del método 1, para asegurarse de que comprende los fundamentos de tales cálculos. La comprensión es siempre mucho mejor que la aplicación a ciegas de las fórmulas.

Método 1: Uso de la fórmula de probabilidad binomial En una distribución binomial, las probabilidades pueden calcularse con el uso de la fórmula de la probabilidad binomial.

Fórmula 4-5 $$P(x) = \frac{n!}{(n-x)!x!} \cdot p^x \cdot q^{n-x} \qquad \text{para } x = 0, 1, 2, \ldots, n$$

donde n = número de ensayos
x = número de éxitos en n ensayos
p = probabilidad de éxito en cualquier ensayo
q = probabilidad de fracaso en cualquier ensayo ($q = 1 - p$)

El símbolo de factorial !, que se introdujo en la sección 3-7, denota el producto de factores decrecientes. Dos ejemplos de factoriales son $3! = 3 \cdot 2 \cdot 1 = 6$ y $0! = 1$ (por definición). Muchas calculadoras incluyen una tecla para el factorial, al igual que una tecla con ${}_nC_r$ que permite simplificar los cálculos. Para las calculadoras con esa tecla, utilice esta versión de la fórmula de probabilidad binomial (donde n, x, p y q son iguales en la fórmula 4-5):

$$P(x) = {}_nC_r \cdot p^x \cdot q^{n-x}$$

La calculadora TI-83 Plus se diseñó para calcular de manera automática las probabilidades binomiales por medio de tal fórmula. Los detalles para el manejo de la calculadora TI-83 Plus se explicarán más adelante en esta sección.

EJEMPLO Análisis de respuestas de opción múltiple Aplique la fórmula de la probabilidad binomial para calcular la probabilidad de tener exactamente tres respuestas correctas, cuando se adivina al azar en las cuatro preguntas de opción múltiple. Esto es, calcule $P(3)$, siendo que $n = 4$, $x = 3$, $p = 0.2$ y $q = 0.8$.

SOLUCIÓN Con los valores dados de n, x, p y q en la fórmula de probabilidad binomial (fórmula 4-5), obtenemos

$$\begin{aligned} P(x) &= \frac{4!}{(4-3)!3!} \cdot 0.2^3 \cdot 0.8^{4-3} \\ &= \frac{4!}{1!3!} \cdot 0.008 \cdot 0.8 \\ &= (4)(0.008)(0.8) = 0.0256 \end{aligned}$$

La probabilidad de tener exactamente tres respuestas correctas de cuatro, es de 0.0256.

Sugerencia para el cálculo: Cuando se calcula una probabilidad con la fórmula de probabilidad binomial, es útil obtener un solo número para $n!/(n-x)!x!$, un solo número para p^x y un solo número para q^{n-x}, así como luego simplemente multiplicar los tres factores, como se muestra al final de los cálculos del ejemplo anterior. No redondee demasiado al calcular esos tres factores; hágalo al final.

Método 2: Uso de la tabla A-1 del Apéndice A En algunos casos calculamos fácilmente las probabilidades binomiales con sólo remitirnos a la tabla A-1 del Apéndice A. Primero localice n y el valor correspondiente de x deseado. En esta etapa, debe aislarse un renglón de números. Ahora, alinee dicho renglón con la probabilidad correspondiente de p, usando la columna que cruza. El número representa la probabilidad deseada. El 0+ indica una probabilidad tan pequeña como 0.000000345.

Al margen se muestra parte de la tabla A-1. Con $n = 4$ y $p = 0.2$ en una distribución binomial, las probabilidades de 0, 1, 2, 3 y 4 éxitos son de 0.410, 0.410, 0.154, 0.026 y 0.002, respectivamente.

De la tabla A-1:

n	x	p 0.20
4	0	0.410
	1	0.410
	2	0.154
	3	0.026
	4	0.002

↓

Distribución de probabilidad binomial para $n = 4$ y $p = 0.2$

x	$P(x)$
0	0.410
1	0.410
2	0.154
3	0.026
4	0.002

EJEMPLO Use la parte de la tabla A-1 (para $n = 4$ y $p = 0.2$), que está al margen, para calcular lo siguiente:

a. La probabilidad de exactamente tres éxitos.

b. La probabilidad de *al menos* tres éxitos.

SOLUCIÓN

a. En la tabla A-1 se observa que cuando $n = 4$ y $p = 0.2$, la probabilidad de $x = 3$ está dada por $P(3) = 0.026$, que es el mismo valor (excepto por el redondeo) que se calcula con la fórmula de probabilidad binomial en el ejemplo anterior.

b. "Al menos" tres éxitos, significa que el número de éxitos es 3 o 4.

$$\begin{aligned} P(\text{al menos } 3) &= P(3 \text{ o } 4) \\ &= P(3) + P(4) \\ &= 0.026 + 0.002 \\ &= 0.028 \end{aligned}$$

En el inciso *b* de la solución anterior, si calculásemos P (al menos 3) por medio de la fórmula de probabilidad binomial, necesitaríamos aplicar la fórmula en dos ocasiones para calcular dos probabilidades diferentes, que después debían sumarse. Al elegir entre la fórmula y la tabla, es más lógico el uso de esta última. Desafortunadamente, la tabla A-1 incluye sólo un número limitado de valores de n y de p, por lo que no siempre sirve; entonces, habrá que calcular las probabilidades con la fórmula de probabilidad binomial, con un programa de cómputo o una calculadora, como en el siguiente método.

Método 3: Uso de las herramientas tecnológicas El STATDISK, Minitab, Excel y la calculadora TI-83 PLus resultan útiles para calcular probabilidades binomiales. Al final de esta sección, presentaremos los detalles para el uso de estas herramientas tecnológicas. Por ahora, observe las representaciones visuales comunes que listan las probabilidades binomiales para $n = 4$ y $p = 0.2$.

STATDISK

Binomial Probabilities

Num Trials, n: 4
Success Prob, p: 0.2

Evaluate | Help | Plot

x	P(x)	P(x≤)	P(x≥)
0	0.40960	0.40960	1.00000
1	0.40960	0.81920	0.59040
2	0.15360	0.97280	0.18080
3	0.02560	0.99840	0.02720
4	0.00160	1.00000	0.00160

Mean	0.800000
St Dev	0.800000
Variance	0.640000

Minitab

```
Binomial with n = 4 and p = 0.200000

     x        P( X = x )
  0.00          0.4096
  1.00          0.4096
  2.00          0.1536
  3.00          0.0256
  4.00          0.0016
```

Excel

	A	B
1	0	0.4096
2	1	0.4096
3	2	0.1536
4	3	0.0256
5	4	0.0016

TI-83 Plus

L1	L2	L3 2
0	.4096	------
1	.4096	
2	.1536	
3	.0256	
4	.0016	

L2(6) =

Ya que ahora conocemos tres métodos diferentes para calcular las probabilidades binomiales, he aquí una estrategia efectiva y eficiente:

1. Utilice un programa de cómputo o una calculadora TI-83 Plus, si están disponibles.
2. Si no dispone de un programa de cómputo ni de la calculadora TI-83 Plus, utilice la tabla A-1.
3. Si no dispone de un programa de cómputo ni de una calculadora, y no le es posible hallar las probabilidades en la tabla A-1, entonces utilice la fórmula de probabilidad binomial.

Fundamentos de la fórmula de probabilidad binomial

La fórmula de probabilidad binomial es la base de los tres métodos que se presentan en esta sección. En lugar de aceptarse y usar la fórmula ciegamente, veamos cómo funciona.

En esta sección ya utilizamos la fórmula de probabilidad binomial para calcular la probabilidad de obtener exactamente tres respuestas correctas, cuando se hacen conjeturas al azar en las preguntas con cuatro opciones. Para cada pregunta

Voltaire gana la lotería

En 1729, el filósofo *Voltaire* se hizo rico al diseñar un esquema para ganar en la lotería de París. El gobierno organizó una lotería para reembolsar bonos municipales que habían perdido cierto valor. La ciudad aportó grandes cantidades de dinero, con el efecto neto de que el valor total de los premios fuese mayor que el costo de todos los boletos. *Voltaire* organizó un grupo y compró todos los boletos de la lotería mensual y ganó durante más de un año. Un participante de la lotería del estado de Nueva York trató de ganar una parte de un premio excepcionalmente grande, que había crecido gracias a la falta de ganadores previos. Él quería extender un cheque por $6,135,756, que cubriría todas las combinaciones, pero el estado se rehusó y afirmó que esto cambiaría la naturaleza de la lotería.

hay cinco respuestas posibles, de modo que la probabilidad de una respuesta correcta es $1/5$ o 0.2. Si utilizamos la regla de la multiplicación, de la sección 3-4, obtenemos el siguiente resultado:

$$\begin{aligned} &P(\text{3 respuestas correctas, seguidas por 1 respuesta incorrecta}) \\ &= 0.2 \cdot 0.2 \cdot 0.2 \cdot 0.8 \\ &= 0.2^3 \cdot 0.8 \\ &= 0.0064 \end{aligned}$$

Este resultado es incorrecto porque supone que las *primeras* respuestas son correctas y que la *última* es incorrecta, pero hay otros acomodos posibles para tres respuestas correctas y una respuesta incorrecta.

En la sección 3-7, vimos que con tres elementos idénticos (como respuestas *correctas*) y otro elemento (como una respuesta *incorrecta*) el número total de acomodos (permutaciones) es $4!/(4-3)!3!$ o 4. Cada uno de estos distintos acomodos tiene una probabilidad de $0.2^3 \cdot 0.8$, de modo que la probabilidad total es la siguiente:

$$P(\text{3 correctas de 4}) = \frac{4!}{(4-3)!3!} \cdot 0.2^3 \cdot 0.8^1$$

Generalice los resultados como sigue: reemplace 4 por n, reemplace x por 3, reemplace 0.2 por p, reemplace 0.8 por q, y exprese el componente de 1 como $4 - 3$, que puede ser reemplazado por $n - x$. El resultado es la fórmula de probabilidad binomial, es decir, la fórmula es una combinación de la regla de la multiplicación de probabilidad y la regla de conteo para el número de acomodos de n elementos, cuando x de ellos son idénticos entre sí, y los otros $n - x$ son idénticos entre sí. (Véanse los ejercicios 9 y 10).

$$P(x) = \overbrace{\frac{n!}{(n-x)!x!}}^{\text{Número de resultados con exactamente } x \text{ éxitos en } n \text{ ensayos}} \cdot \overbrace{p^x \cdot q^{n-x}}^{\text{Probabilidad de } x \text{ éxitos en } n \text{ ensayos, para cualquier orden}}$$

Encuestas sensibles

En ocasiones, las personas que contestan las encuestas se rehúsan a responder con honestidad preguntas sobre un tema sensible como el robo al empleador o el sexo. Stanley Warner (York University, Ontario) diseñó un sistema que produce resultados más precisos en casos como éstos. Como ejemplo, pregunte a algunos empleados si cometieron un robo durante el año anterior, luego pídales que lancen una moneda. Instrúyalos para que respondan que no en caso de que no hayan robado y la moneda caiga en cara. Si no es así, deben responder que sí. Los empleados suelen ser más honestos porque el lanzamiento de la moneda los ayuda a proteger su privacidad. Después, es posible utilizar la teoría de la probabilidad para analizar las respuestas, de modo que se obtengan resultados más precisos.

Utilizando la tecnología

El método 3 de esta sección incluyó el uso del STATDISK, Minitab, Excel o una calculadora TI-83 Plus. Al aplicar los siguientes procedimientos para el cálculo de probabilidades binomiales, aparecen las representaciones visuales del uso del método 3.

STATDISK Seleccione **Analysis** del menú principal; después, seleccione la opción **Binomial Probabilities**. Introduzca los valores requeridos de n y p, aparecerá la distribución de probabilidad completa. Las otras columnas representan las probabilidades acumulativas que se obtienen al sumar los valores de $P(x)$, conforme sube o baja, a lo largo de la columna.

Minitab Primero introduzca la columna C1 de los valores x de los que desea conocer las probabilidades (tales como 0, 1, 2, 3, 4); después, seleccione **Calc** del menú principal y proceda a elegir los siguientes elementos: **Distribuciones de probabilidad** y **Binomial.** Seleccione **Probabilities**, introduzca el número de ensayos, la probabilidad de éxito y C1 en la columna de entrada; después, haga clic en **OK**.

Excel Liste los valores de x en la columna A (tales como 0, 1, 2, 3, 4). Haga clic en la celda B1, en f_x de la barra de herramientas y seleccione la categoría de función **Statistical** y luego **BINOMDIST**. En el cuadro de diálogo, introduzca A1 para el número de éxitos, el número de ensayos, la probabilidad y 0 para la distribución binomial (en lugar de 1 para la distribución binomial acumulativa). Debe aparecer un valor en la celda B1. Haga clic y lleve la esquina derecha inferior de la celda B1 hacia abajo de la columna para emparejarla con los datos de la columna A; después, libere el botón del ratón.

TI-83 Plus Presione **2nd VARS** (para obtener **DISTR**, que denota "distribuciones"); después, seleccione la opción identificada por **binompdf(.**. Complete **binompdf (n, p, x)** con los valores específicos de n, p y x; luego, presione **ENTER;** el resultado será la probabilidad de obtener x éxitos en n ensayos.

También es posible elegir **binompdf (n, p)** para obtener una lista de *todas* las probabilidades correspondientes a $x = 1, 2, \ldots, n$. Puede almacenar esta lista en L2 si presiona **STO→L2**. Después, podría introducir los valores de 0, 1, 2, . . . , n en la lista L1, lo cual le permitiría calcular estadísticos (con **STAT, CALC** y luego **L1, L2**) o ver la distribución en formato de tabla (presionando **STAT** y luego **EDIT**).

4-3 Destrezas y conceptos básicos

Identificación de distribuciones normales. *En los ejercicios 1 a 8 determine si el procedimiento dado resulta en una distribución binomial. En aquellos que no sean binomiales, identifique al menos uno de los requisitos que no se satisfacen.*

1. Aplicar una encuesta en la que se le pregunta a las personas lo que piensan del presidente actual.

2. Aplicar una encuesta a 1012 sujetos y registrar si hay una respuesta "no debe" a la siguiente pregunta: "¿La clonación de seres humanos debe o no permitirse?"

3. Tirar un dado balanceado 50 veces.

4. Tirar un dado cargado 50 veces y calcular el número de veces que resulta 5.

5. Registrar el género de 250 bebés recién nacidos.

6. Determinar si cada uno de 3000 marcapasos cardiacos es aceptable o defectuoso.

7. Girar una ruleta 12 veces.

8. Girar una ruleta 12 veces y calcular el número de ocasiones en que el resultado es un número impar.

9. Cálculo de probabilidades con respuestas de adivinación Cada pregunta de opción múltiple tiene cinco posibles respuestas, una de las cuales es la correcta. Suponga que adivina las respuestas de tres de estas preguntas.

continúa

a. Utilice la regla de la multiplicación para calcular la probabilidad de que las dos primeras conjeturas sean incorrectas y que la tercera sea correcta. Es decir, calcule P (IIC), donde C denota una respuesta correcta e I una incorrecta.
b. Inicie con IIC y haga una lista completa de los distintos acomodos posibles de dos respuestas incorrectas y una correcta; después, calcule la probabilidad de cada dato en la lista.
c. Con base en los resultados anteriores, ¿cuál es la probabilidad de tener exactamente una respuesta correcta cuando se hacen tres adivinaciones?

10. Cálculo de probabilidades con respuestas de adivinación Un examen consta de preguntas de opción múltiple con cuatro respuestas posibles, una de las cuales es la correcta. Suponga que adivina las respuestas a seis de estas preguntas.
a. Utilice la regla de la multiplicación para calcular la probabilidad de que las dos primeras conjeturas sean incorrectas y que las cuatro últimas sean correctas. Es decir, calcule P (IICCCC), donde C denota una respuesta correcta e I una incorrecta.
b. Inicie con IICCCC y haga una lista completa de los distintos acomodos posibles de dos respuestas incorrectas y cuatro correctas; después, calcule la probabilidad de cada dato en la lista.
c. Con base en los resultados anteriores, ¿cuál es la probabilidad de tener exactamente cuatro respuestas correctas cuando se hacen seis adivinaciones?

Uso de la tabla A-1. *En los ejercicios 11 a 16 suponga que un procedimiento produce una distribución binomial, con un ensayo repetido* n *veces. Utilice la tabla A-1 para calcular la probabilidad de* x *éxitos, dada la probabilidad* p *de éxito en un solo ensayo.*

11. $n = 2, x = 0, p = 0.01$
12. $n = 7, x = 2, p = 0.01$
13. $n = 4, x = 3, p = 0.95$
14. $n = 6, x = 5, p = 0.99$
15. $n = 10, x = 4, p = 0.95$
16. $n = 11, x = 7, p = 0.05$

Uso de la fórmula de probabilidad binomial. *En los ejercicios 17 a 20 suponga que un procedimiento produce una distribución binomial, con un ensayo que se repite* n *veces. Utilice la fórmula de probabilidad binomial para calcular la probabilidad de* x *éxitos, dada la probabilidad* p *de éxito en un solo ensayo.*

17. $n = 6, x = 4, p = 0.55$
18. $n = 6, x = 2, p = 0.45$
19. $n = 8, x = 3, p = 1/4$
20. $n = 10, x = 8, p = 1/3$

Uso de resultados de computadora. *En los ejercicios 21 a 14 remítase a la representación visual de Minitab al margen. Las probabilidades se obtuvieron al introducir los valores de $n = 6$ y $p = 0.723$. Hay una probabilidad de 0.723 de que un vuelo de American Airlines, que se selecciona aleatoriamente, llegue a tiempo (de acuerdo con datos del Departamento del Transporte de Estados Unidos). En cada caso, suponga que se seleccionan aleatoriamente seis vuelos de American Airlines y calcule la probabilidad indicada.*

Binomial con n = 6 y p = 0.723000

x	$P(X = x)$
0.00	0.0005
1.00	0.0071
2.00	0.0462
3.00	0.1607
4.00	0.3145
5.00	0.3283
6.00	0.1428

21. Calcule la probabilidad de que al menos cinco vuelos de American Airlines lleguen a tiempo. ¿Es poco común que al menos cinco de seis vuelos de American Airlines lleguen a tiempo?

22. Calcule la probabilidad de que a lo sumo dos vuelos de American Airlines lleguen a tiempo. ¿Es poco común que a lo sumo dos de seis vuelos de American Airlines lleguen a tiempo?

23. Calcule la probabilidad de que más de un vuelo de American Airlines llegue a tiempo. ¿Es poco común que *no* más de uno de seis vuelos de American Airlines llegue a tiempo?

24. Calcule la probabilidad de que al menos un vuelo de American Airlines llegue a tiempo. ¿Es poco común que *no* haya al menos uno de seis vuelos de American Airlines que llegue a tiempo?

25. Ceguera al color El 9% de los hombres y el 0.25% de las mujeres no pueden distinguir entre los colores rojo y verde. Este tipo de problema visual causa dificultades con las señales de tránsito. Si se seleccionan seis hombres aleatoriamente para un estudio de la percepción de las señales de tránsito, calcule la probabilidad de que exactamente dos de ellos no distingan entre el rojo y el verde.

26. Muestreo de aceptación La compañía Telektronic compra grandes embarques de focos fluorescentes y usa el siguiente plan de muestreo de aceptación: seleccionar aleatoriamente y probar 24 focos; después, aceptar el grupo completo sólo si hay uno o ninguno que no funcione. Si un embarque particular de miles de focos tiene en realidad una tasa de defectos del 4%, ¿cuál es la probabilidad de que el embarque completo se acepte?

27. Auditorías de la IRS La Hemingway Financial Company prepara devoluciones de impuestos para individuos. (Su lema: "También escribimos grandiosas novelas de ficción"). Según el Internal Revenue Service, los individuos que ganan entre 425,000 y 50,000 dólares se auditan en una proporción del 1%. La Hemingway Company prepara cinco devoluciones de impuestos para individuos que están en esa categoría de impuestos, en tanto se audita a tres de ellos.

a. Calcule la probabilidad de que, cuando se seleccione aleatoriamente a cinco personas que ganan entre $25,000 y $50,000, se audite exactamente a tres de ellos.

b. Calcule la probabilidad de que se audite al menos a tres.

c. Con base en los resultados anteriores, ¿qué se concluye acerca de los clientes de Hemingway? ¿Sólo son desafortunados o están siendo blanco de las auditorías?

28. Asistencia del directorio telefónico Un artículo de *USA Today* afirma que "encuestas internas que son pagadas por proveedores de asistencia del directorio telefónico, muestran que incluso las compañías más precisas usan los números incorrectos el 15% de las veces". Suponga que prueba a un proveedor de éstos haciendo 10 solicitudes y también que el proveedor le da números telefónicos incorrectos el 15% de las veces.

a. Calcule la probabilidad de obtener un número incorrecto.

b. Calcule la probabilidad de obtener a lo sumo un número incorrecto.

c. Si usted obtiene a lo sumo un número incorrecto, ¿parecería que la tasa de números incorrectos no es del 15%, como se afirma?

29. Vuelos sobresaturados Air America tiene la política de registrar a 15 personas en un avión donde sólo caben 14. (Estudios anteriores revelaron que sólo el 85% de los pasajeros que se registran usan el vuelo). Calcule la probabilidad de que, si Air America registra a 15 personas, no haya suficientes asientos disponibles. ¿Será la probabilidad suficientemente baja, de modo que la sobreventa no sea un problema real para los pasajeros?

30. Reacción al fármaco En una prueba clínica del fármaco Viagra, se encontró que el 4% de los individuos en el grupo placebo sufrieron dolores de cabeza.

a. Suponiendo que la misma tasa del 4% se aplica a quienes toman Viagra, calcule la probabilidad de que, entre ocho usuarios del Viagra, tres experimenten dolores de cabeza.

b. Suponiendo que la misma tasa del 4% se aplica a quienes toman Viagra, calcule la probabilidad de que, entre ocho usuarios de Viagra que se seleccionan aleatoriamente, todos ellos experimenten dolores de cabeza.

c. Si los ocho usuarios de Viagra experimentaran dolores de cabeza, ¿parecería que la tasa de dolores de cabeza de los usuarios de Viagra es diferente de la tasa del 4% de los sujetos del grupo de placebo? Explique.

31. Encuestas a televidentes El programa de televisión *60 minutos*, de la CBS, ha logrado éxito por muchos años. Recientemente registró una audiencia de 20, lo que significa que de todos los televisores en uso el 20% se sintonizan en *60 minutos* (según datos de Nielsen Media Research). Suponga que un anunciante desea verificar dicho valor del 20%, realizando su propia encuesta, y que inicia una encuesta piloto con 10 hogares que tienen la televisión encendida en el momento en que se transmite el programa *60 minutos*.

a. Calcule la probabilidad de que ninguno de los hogares esté sintonizando *60 minutos*.

b. Calcule la probabilidad de que al menos uno de los hogares esté sintonizando *60 minutos*.

c. Calcule la probabilidad de que a lo sumo uno de los hogares esté sintonizando *60 minutos*.

d. Si a lo sumo un hogar está sintonizando *60 minutos*, ¿será incorrecto el valor de una audiencia del 20%? ¿Por qué?

32. Programas de acción afirmativa Se realizó un estudio para determinar si había diferencias significativas entre estudiantes de medicina que se aceptaron por medio de programas especiales (como el de acción afirmativa) y estudiantes de medicina que se aceptaron a través de los criterios regulares de admisión. Se encontró que el 94% de los estudiantes de medicina que se aceptaron a través de programas especiales se graduaron (según datos del *Journal of the American Medical Association).*

a. Si se seleccionan aleatoriamente 10 de los estudiantes de los programas especiales, calcule la probabilidad de que al menos nueve se gradúen.

b. ¿Sería poco común que de 10 estudiantes de los programas especiales, que se seleccionaron aleatoriamente, sólo se graduaran siete? ¿Por qué?

33. Identificación de la discriminación por género Después de que la rechazaran para un empleo, Kim Kelly se entera de que la Bellevue Advertising Company sólo contrató a dos mujeres entre los últimos 20 empleados nuevos. También, de que el grupo de solicitantes es muy grande, y que incluye un número aproximadamente igual de hombres y mujeres calificados. Ayúdele a presentar cargos por discriminación por género, calculando la probabilidad de que dos o menos mujeres se incluyan en una contratación de 20 personas, suponiendo que no hay discriminación que se basa en el género. ¿Apoya la probabilidad resultante esos cargos?

34. Máquina tragamonedas del autor El autor compró una máquina tragamonedas que se configuró de tal forma que hay una probabilidad de $1/2000$ de ganarse el premio mayor en cualquier ensayo individual. Aun cuando nadie consideraría seriamente hacer trampa al autor, suponga que un invitado afirma haber jugado con la máquina cinco veces y ganado en dos ocasiones.

a. Calcule la probabilidad de exactamente dos premios en cinco ensayos.

b. Calcule la probabilidad de al menos dos premios en cinco ensayos.

c. ¿Parece válida la afirmación del invitado de dos triunfos en cinco juegos? Explique.

35. Prueba de la eficacia de la técnica de selección del género El problema del capítulo describe la distribución de probabilidad del número de niñas *x* resultantes cuando se seleccionan aleatoriamente 14 bebés recién nacidos. Suponga que otro experimento clínico incluye 12 bebés recién nacidos. Utilice el mismo formato de la tabla 4-1 y construya una tabla para la distribución de probabilidad que resulta de los 12 nacimientos; después, determine si una técnica de selección del género sería efectiva si nacen nueve niñas y tres niños.

36. Cursos de posgrado El Market Research Institute encontró que de los individuos que se graduaron de la universidad desde hace al menos 10 años, con empleo y edades entre 30 y 55 años, el 57% tomaron cursos universitarios tras haberse graduado (según el *USA Today*). Si usted selecciona aleatoriamente a cinco individuos que se graduaron de la universidad desde hace al menos 10 años, con edades entre 30 y 55 años, y descubre que sólo uno de ellos tomó cursos, ¿debe pensar que la tasa del 57% es incorrecta? Explique.

4-3 Más allá de lo básico

37. Si un procedimiento cumple con todas las condiciones de una distribución binomial, excepto que el número de ensayos no es fijo, entonces se puede utilizar una **distribución geométrica**. La probabilidad de obtener el primer éxito en el ensayo x-ésimo está dada por $P(x) = p(1 - p)^{x-1}$, donde p es la probabilidad de éxito en cualquier ensayo. Suponga que la probabilidad de un componente de computadora defectuoso es de 0.2. Calcule la probabilidad de que el primer defecto se descubra en el séptimo componente que se probó.

38. Si realizamos un muestreo sin reemplazo de una población finita pequeña, no debe usarse la distribución binomial porque los sucesos no son independientes. Si el muestreo se hace sin reemplazo y los resultados pertenecen a uno de dos tipos, podemos usar la **distribución hipergeométrica**. Si una población tiene A objetos de un tipo, mientras que los objetos B restantes son de otro tipo, y si se muestrean sin reemplazo n objetos, entonces la probabilidad de obtener x objetos del tipo A y $n - x$ objetos del tipo B es

$$P(x) = \frac{A!}{(A - x)!x!} \cdot \frac{B!}{(B - n + x)!(n - x)!} \div \frac{(A + B)!}{(A + B - n)!n!}$$

En la lotería 54, un participante selecciona seis números del 1 al 54 (sin repetición); después, se selecciona aleatoriamente una combinación de seis números ganadores. Calcule la probabilidad de obtener:

a. Los seis números ganadores.
b. Exactamente cinco de los números ganadores.
c. Exactamente tres de los números ganadores.
d. Ningún número ganador.

39. La distribución binomial se aplica sólo a casos que impliquen dos tipos de resultados, mientras que la **distribución multinomial** supone más de dos categorías. Suponga que tenemos tres tipos de resultados mutuamente excluyentes, que se denotan por A, B y C. Sean $P(A) = p_1$, $P(B) = p_2$, y $P(C) = p_3$. En n ensayos independientes, la probabilidad de x_1 resultados tipo A, x_2 resultados tipo B y x_3 resultados tipo C está dada por:

$$\frac{n!}{(x_1!)(x_2!)(x_3!)} \cdot p_1^{x_1} \cdot p_2^{x_2} \cdot p_3^{x_3}$$

Un experimento en genética incluye seis genotipos mutuamente excluyentes identificados como A, B, C, D, E y F, todos igual de probables. Si se prueba a 20 descendientes, calcule la probabilidad de obtener con exactitud cinco A, cuatro B, tres C, dos D, tres E y tres F, al expandir la expresión anterior, de modo que se aplique a seis tipos de resultados y no sólo a tres.

4-4 Media, varianza y desviación estándar para la distribución binomial

En el capítulo 2, vimos que cuando se investigan conjuntos de datos reales, las siguientes características suelen ser muy importantes: **1.** las medidas de tendencia central, **2.** las medidas de variación, **3.** la naturaleza de la distribución, **4.** la presencia de datos distantes y **5.** un patrón a lo largo del tiempo. (Recuerde que utilizamos “CVDDT” como herramienta para evocar dichas características). Un aspecto central de este capítulo es que las distribuciones de probabilidad describen

¿Prevalecen los niños o las niñas en las familias?

El autor de este libro, sus hermanos y sus sobrinos suman un total de 11 hombres y sólo una mujer. ¿Será este ejemplo un fenómeno en el que un género particular prevalece en una familia? El tema se estudió examinando una muestra aleatoria de 8770 hogares de Estados Unidos. Los resultados se reportaron en la revista *Chance*, en el artículo "Does Having Boys or Girls run in the Family?", escrito por Joseph Rodgers y Debby Doughty. Parte de su análisis implica el uso de la distribución de probabilidad binomial que se estudia en esta sección. Ellos concluyeron que "no encontramos evidencias contundentes de que un sexo prevalezca más en la familia".

lo que *probablemente* sucederá, y no lo que en realidad sucedió. En la sección 4-2 estudiamos métodos para analizar las distribuciones de probabilidad a través del cálculo de la media, la desviación estándar y el histograma de probabilidad. Como una distribución binomial es un tipo especial de distribución de probabilidad, utilizamos las fórmulas 4-1, 4-3 y 4-4 (de la sección 4-2) para calcular la media, la varianza y la desviación estándar. Por fortuna, es posible simplificar mucho tales fórmulas para las distribuciones binomiales, como se muestra a continuación.

Para cualquier distribución de probabilidad discreta		**Para distribuciones binomiales**	
Fórmula 4-1	$\mu = \Sigma[x \cdot P(x)]$	**Fórmula 4-6**	$\mu = np$
Fórmula 4-3	$\sigma^2 = \Sigma[x^2 \cdot P(x)] - \mu^2$	**Fórmula 4-7**	$\sigma^2 = npq$
Fórmula 4-4	$\sigma = \sqrt{\Sigma[x^2 \cdot P(x)] - \mu^2}$	**Fórmula 4-8**	$\sigma = \sqrt{npq}$

EJEMPLO Género de los hijos En la sección 4-2 incluimos un ejemplo para ilustrar los cálculos de μ y σ. Utilizamos el ejemplo de la variable aleatoria x que representa el número de niñas entre 14 nacimientos. (Véase la tabla 4-2 en la página 185, con los cálculos que ejemplifican las fórmulas 4-1 y 4-4.) Aplique las fórmulas 4-6 y 4-8 para calcular la media y la desviación estándar del número de niñas en grupos de 14 nacimientos.

SOLUCIÓN Con los valores $n = 14$, $p = 0.5$ y $q = 0.5$, las fórmulas 4-6 y 4-8 se aplican como sigue:

$$\mu = np = (14)(0.5) = 7.0$$
$$\sigma = \sqrt{npq}$$
$$= \sqrt{(14)(0.5)(0.5)} = 1.9 \quad \text{(redondeado)}$$

Si compara estos cálculos con los que se requieren en la tabla 4-3, resultará evidente que las fórmulas 4-6 y 4-8 son mucho más fáciles de usar.

La fórmula 4-6 para la media es intuitivamente lógica. Si le preguntáramos a cualquier estudiante de estadística cuántas niñas se esperaría en 100 nacimientos, la respuesta común sería 50, que puede generalizarse fácilmente como $\mu = np$. La varianza y la desviación estándar no se justifican tan fácilmente, por lo que omitiremos los complicados manejos algebraicos que conducen a las fórmulas 4-7 y 4-8. En su lugar, remítase de nuevo al ejemplo anterior y a la tabla 4-3 para verificar que, en una distribución binomial, las fórmulas 4-6, 4-7 y 4-8 producirán los mismos resultados que las fórmulas 4-1, 4-3 y 4-4.

EJEMPLO Selección del género El problema del capítulo incluyó un ensayo preliminar con 14 parejas que deseaban tener niñas. Aun cuando el resultado de 13 niñas en 14 nacimientos parece indicar que el método MicroSort es eficaz para la selección del género, tendríamos mucho mayor confianza en dicha conclusión si el tamaño de la muestra fuese considerablemente mayor que 14. Suponga que el método MicroSort se utiliza con 100 parejas que tendrán un bebé. También, que el resultado es de 68 niñas entre 100 bebés.

a. Suponiendo que el método de selección del género MicroSort no produce efecto alguno, calcule la media y la desviación estándar del número de niñas en grupos de 100 bebés que se seleccionaron aleatoriamente.

b. *Interprete* los valores del inciso *a* para determinar si estos resultados (68 niñas entre 100 bebés) apoyan la afirmación de que el método MicroSort de selección del género es efectivo.

SOLUCIÓN

a. Suponiendo que el método MicroSort no produce efecto alguno, así como que las niñas y los niños son igualmente probables, tenemos $n = 100$, $p = 0.5$ y $q = 0.5$. Podemos calcular la media y la desviación estándar con el uso de las fórmulas 4-6 y 4-8 de la siguiente manera:

$$\mu = np = (100)(0.5) = 50$$
$$\sigma = \sqrt{npq} = \sqrt{(100)(0.5)(0.5)} = 5$$

Para grupos de 100 parejas con un bebé, el número medio de niñas es de 50 y la desviación estándar es de 5.

b. Ahora debemos interpretar los resultados para determinar si 68 niñas, entre 100 bebés, implica algo que podría ocurrir fácilmente por azar, o si es tan improbable que el método MicroSort de selección del género parece ser efectivo. Utilizaremos la regla práctica del intervalo de la siguiente manera:

$$\text{valor máximo común: } \mu + 2\sigma = 50 + 2(5) = 60$$
$$\text{valor mínimo común: } \mu - 2\sigma = 50 - 2(5) = 40$$

INTERPRETACIÓN De acuerdo con la regla práctica del intervalo, los valores se consideran comunes si están entre 40 y 60, de modo que 68 niñas es un resultado poco común, ya que no se encuentra entre 40 y 60. Es muy poco probable que resulten 68 niñas en 100 nacimientos sólo por el azar. Si en realidad obtuviésemos 68 niñas entre 100 nacimientos, deberíamos buscar una explicación alternativa a la del azar. Si las 100 parejas utilizaron el método MicroSort de selección del género, parecería que es efectivo para incrementar la posibilidad de que un hijo sea niña.

Usted debe desarrollar la habilidad para calcular medias y desviaciones estándar con el uso de las fórmulas 4-6 y 4-8, pero es especialmente importante aprender a *interpretar* los resultados con el empleo de esos valores. La regla práctica del intervalo, como se ilustra en el inciso *b* del ejemplo anterior, sugiere que los valores son poco comunes si caen fuera de los siguientes límites:

$$\textbf{valor máximo común} = \boldsymbol{\mu + 2\sigma}$$
$$\textbf{valor mínimo común} = \boldsymbol{\mu - 2\sigma}$$

4-4 Destrezas y conceptos básicos

Cálculo de μ, σ y valores poco comunes. *En los ejercicios 1 a 4 suponga que un procedimiento produce una distribución binomial con* n *ensayos, y que la probabilidad de éxito de un ensayo es* p*. Utilice los valores de* n *y* p *dados para calcular la media* μ *y la des-*

Los estados controlan las selecciones de lotería

Muchos estados permiten una lotería en la que los jugadores seleccionan cuatro dígitos, como 1127 (el cumpleaños del autor). Si un jugador paga un dólar y selecciona la secuencia ganadora en el orden correcto, gana un premio de $5000. Los estados observan las selecciones de números y, si una secuencia en particular se elige con demasiada frecuencia, se prohíbe a los jugadores hacer más apuestas con tal secuencia. Las máquinas de lotería se controlan de tal manera que, una vez que una secuencia popular alcanza cierto nivel de ventas, no aceptarán más dicha secuencia. Lo anterior evita que los estados paguen más de lo que reciben. Los críticos afirman que la práctica es injusta. Según William Thompson, un experto en apuestas de la Universidad de Nevada en Las Vegas, "ellos (los estados) afirman que quieren estar en el negocio del juego, pero no desean ser jugadores. Esto hace una farsa del juego de números".

viación estándar σ. Además, use la regla práctica del intervalo para calcular el valor mínimo común $\mu - 2\sigma$ y el valor máximo común $\mu + 2\sigma$.

1. $n = 400, p = 0.2$

2. $n = 250, p = 0.45$

3. $n = 1984, p = 3/4$

4. $n = 767, p = 1/6$

5. Respuestas de adivinación Varios estudiantes no se prepararon para enfrentar un examen sorpresa de verdadero/falso de 10 preguntas, por lo cual todas sus respuestas son conjeturas.
a. Calcule la media y la desviación estándar del número de respuestas correctas de esos estudiantes.
b. ¿Sería poco común que un estudiante pasara el examen adivinando y que obtuviera al menos siete respuestas correctas? ¿Por qué?

6. Respuestas de adivinación Varios estudiantes no se prepararon para presentar un examen de opción múltiple de 10 preguntas, por lo cual todas sus respuestas son conjeturas. Cada pregunta tiene cinco respuestas posibles, pero sólo una de ellas es correcta.
a. Calcule la media y la desviación estándar del número de respuestas correctas de esos estudiantes.
b. ¿Sería poco común que un estudiante pasara el examen adivinando y que obtuviera al menos siete respuestas correctas? ¿Por qué?

7. Juego de ruleta Si apuesta en cualquier número de la ruleta, su probabilidad de ganar es 1/38. Suponga que apuesta a un solo número en cada uno de 100 giros consecutivos.
a. Calcule la media y la desviación estándar del número de triunfos.
b. ¿Sería poco común no ganar ni una ocasión en los 100 ensayos? ¿Por qué?

8. Personas zurdas El 10% de los adultos estadounidenses son zurdos. Una clase de estadística tiene 25 estudiantes.
a. Calcule la media y la desviación estándar del número de estudiantes zurdos en la clase con 25 estudiantes.
b. ¿Sería poco común hacer una encuesta a una clase con 25 estudiantes y encontrar que 5 de ellos son zurdos? ¿Por qué?

9. Análisis de los resultados de un experimento de selección del género Un experimento, en relación con un método de selección del género, incluye un grupo control de 15 parejas que no reciben ningún tratamiento creado para influir en el género de sus hijos. Cada una de las 15 parejas tiene un hijo.
a. Elabore una tabla que liste los posibles valores de la variable aleatoria x (que representa el número de niñas entre 15 nacimientos) y las probabilidades correspondientes.
b. Calcule la media y la desviación estándar del número de niñas en grupos de 15 nacimientos como éstos.
c. Si las parejas tienen 10 niñas y cinco niños, ¿será poco común este resultado? ¿Por qué?

10. Mensajes descifrados La Central Intelligence Agency tiene especialistas que analizan la secuencia de letras del alfabeto, en un intento por descifrar mensajes que se interceptan. En un texto estándar en inglés, la letra r se utiliza en una proporción del 7.7%.
a. Calcule la media y la desviación estándar del número de veces que la letra r aparecerá en una página común de 2,600 caracteres.
b. En un mensaje que se interceptó que iba hacia Irak, se encontró que en una página con 2,600 caracteres la letra r aparecía 175 veces. ¿Es esto poco común?

11. Determinar si disminuyen las quejas después de un programa de entrenamiento La Newtower Department Store recibió una tasa de quejas de los clientes del 3.2%, e intenta disminuir esta tasa con un programa de entrenamiento para los empleados. Una vez que se completó el programa, se localizó a 850 clientes y se descubrió que sólo siete de ellos se quejaron.

a. Suponiendo que el programa de entrenamiento no tenga efecto alguno, calcule la media y la desviación estándar del número de quejas en grupos de 850 clientes como éste.

b. Con base en los resultados del inciso *a*), ¿es poco común el resultado de siete quejas? ¿Fue efectivo el programa de entrenamiento para disminuir la tasa de quejas?

12. ¿Es azul el 10% de los dulces M&M? Mars, Inc., afirma que el 10% de sus dulces M&M son azules, por lo cual se selecciona aleatoriamente una muestra con 100 de estos dulces.

a. Calcule la media y la desviación estándar del número de dulces azules en grupos de 100 como éste.

b. El conjunto de datos 19, del Apéndice B, consiste en una muestra aleatoria de 100 M&M, de los cuales sólo cinco son azules. ¿Es este resultado poco común? ¿Será incorrecta la tasa del 10%?

13. Teléfonos celulares y cáncer cerebral En un estudio que incluyó 420,000 usuarios de teléfono celular en Dinamarca, se encontró que 135 desarrollaron cáncer cerebral o del sistema nervioso. Si suponemos que este tipo de cáncer no se afecta por los teléfonos celulares, la probabilidad de que una persona adquiera tal enfermedad es de 0.000340.

a. Suponiendo que los teléfonos celulares no se relacionan con el cáncer, calcule la media y la desviación estándar del número de personas, en grupos de 420,000, que pueden tener cáncer cerebral o del sistema nervioso.

b. Con base en los resultados del inciso *a*, ¿será poco común que, entre 420,000 personas, se presenten 135 casos de cáncer cerebral o del sistema nervioso? ¿Por qué?

c. ¿Qué sugieren tales resultados sobre la preocupación pública de que los teléfonos celulares son dañinos para la salud, porque incrementan el riesgo de cáncer cerebral o del sistema nervioso?

14. Fármaco que reduce el colesterol En un ensayo clínico del Lipitor, un fármaco común que se utiliza para disminuir el colesterol, 863 pacientes recibieron un tratamiento de 10 miligramos de tabletas de Atorvastatin. Este grupo incluyó a 19 pacientes que experimentan síntomas de influenza (según datos de Pfizer, Inc.). La probabilidad de que una persona que no recibe tratamiento alguno presente síntomas de influenza es de 0.019.

a. Suponiendo que el Lipitor no produce efectos sobre los síntomas de la influenza, calcule la media y la desviación estándar del número de personas en grupos de 863 individuos que se esperaría presentaran dichos síntomas.

b. Con base en los resultados del inciso *a*, ¿será poco común encontrar que, de 863 personas, 19 experimentan síntomas de influenza? ¿Por qué?

c. Con base en los resultados anteriores, ¿los síntomas de la influenza parecen ser una reacción adversa que debe preocupar a los usuarios de Lipitor?

15. Opiniones sobre la clonación En una encuesta reciente de Gallup se preguntó a 1012 adultos que se seleccionaron aleatoriamente, si "la clonación humana debe o no permitirse". Los resultados mostraron que el 89% de los encuestados opinaron que la clonación no tiene que permitirse.

a. De los 1012 adultos que se encuestaron, ¿cuántos opinaron que no debe permitirse la clonación.

b. Si suponemos que las personas se muestran indiferentes, de manera que el 50% considera que la clonación humana no tiene que permitirse, calcule la media y la desviación estándar del número de personas en grupos de 1012 que se esperaría que opinaron que la clonación no debe permitirse.

c. Con base en los resultados anteriores, ¿parece inusualmente más alto el resultado del 89% de la encuesta de Gallup que la supuesta tasa del 50%? ¿Parece que una inmensa mayoría de adultos opina que la clonación humana no debe permitirse?

16. Choques de automóviles Los conductores de edades entre 20 y 24 años presentan una proporción del 34% de accidentes automovilísticos durante un año (según datos del National Safety Council). Un investigador de una aseguradora descubre que en un grupo de 500 conductores, en el rango de 20 a 24 años, que viven en la ciudad de Nueva York, que se seleccionaron aleatoriamente, el 42% tuvo accidentes el año anterior.

a. ¿Cuántos conductores, en el grupo de 500 individuos de la ciudad de Nueva York, tuvo accidentes el año anterior?

b. Suponiendo que la misma proporción del 34% se aplica a la ciudad de Nueva York, calcule la media y la desviación estándar del número de personas, en grupos de 500 individuos, que se esperaría tuvieran accidentes.

c. Con base en resultados anteriores, ¿parece inusualmente más alto el resultado del 42% de los conductores de la ciudad de Nueva York, comparado con la proporción del 34% de la población general? ¿Se justifican las tasas de seguro más altas para los conductores de la ciudad de Nueva York?

4-4 Más allá de lo básico

17. Uso de la regla empírica y el teorema de Chebyshev Se diseña un experimento para probar la efectividad del método MicroSort de selección del género y se aplica a 100 parejas que desean tener niñas. En un ejemplo que se incluye en esta sección, se utilizó la regla práctica del intervalo para concluir que, de 100 nacimientos, el número de niñas por lo general debe ubicarse entre 40 y 60.

a. La regla empírica (véase sección 2-5) se aplica a distribuciones normales. ¿Es (aproximadamente) normal la distribución de probabilidad binomial para este experimento? ¿Cómo lo sabe?

b. Suponiendo que la distribución sea normal, ¿qué tan probable es que el número de niñas esté entre 40 y 60 (según la regla empírica)?

c. Suponiendo que la distribución es normal, ¿qué tan probable es que el número de niñas esté entre 35 y 65 (según la regla empírica)?

d. De acuerdo con el teorema de Chebyshev, ¿qué concluimos acerca de la probabilidad de que el número de niñas esté entre 40 y 60?

18. Productos aceptables/defectuosos Mario's Pizza Parlor acaba de inaugurarse. Por la falta de entrenamiento de los empleados existe sólo un 0.8 de probabilidad de que una pizza sea comestible. Se acaban de ordenar cinco pizzas. ¿Cuál es el número mínimo de pizzas que deben prepararse para estar al menos 99% seguros de que habrá cinco comestibles?

4-5 La distribución de Poisson

Las dos secciones anteriores se ocuparon de la distribución binomial, que es una de las muchas distribuciones de probabilidad que pueden utilizarse en distintas situaciones. Esta sección introduce la *distribución de Poisson*. Dicha distribución es particularmente importante, ya que con frecuencia se utiliza como modelo matemático para describir comportamientos como la disminución radioactiva, la llegada de pasajeros en una línea y la de aviones a un aeropuerto, los automóviles que van a una gasolinera, los comensales que visitan un restaurante, los estudiantes que asisten a una librería y los usuarios de Internet que se conectan a un sitio. Por ejemplo, suponga que su profesor tiene una hora de asesoría cada lunes a las 11:00, y descubre que durante esa hora el número medio de estudiantes que recibe es de 2.3. Calculamos la probabilidad de que exactamente cuatro estudiantes lleguen en una hora de asesoría, que se selecciona aleatoriamente, un lunes a las 11:00. Utilizamos la distribución de Poisson, que se define de la siguiente manera.

Definición

Distribución de Poisson: distribución de probabilidad discreta que se aplica a las ocurrencias de algún suceso *durante un intervalo específico*. La variable aleatoria x es el número de ocurrencias de un suceso en un intervalo. El intervalo puede ser tiempo, distancia, área, volumen o alguna unidad similar. La probabilidad de que el suceso ocurra x veces, durante un intervalo, está dada por la fórmula 4-9.

Fórmula 4-9 $$P(x) = \frac{\mu^x \cdot e^{-\mu}}{x!} \qquad \text{donde } e \approx 2.71828$$

La distribución de Poisson tiene los siguientes requisitos:

- La variable aleatoria x es el número de ocurrencias de un suceso *durante un intervalo*.
- Las ocurrencias deben ser *aleatorias*.
- Las ocurrencias tienen que ser *independientes* entre sí.
- Las ocurrencias deben estar *uniformemente distribuidas* dentro del intervalo que se emplea.

La distribución de Poisson tiene los siguientes parámetros:

- La media es μ.
- La desviación estándar es $\sigma = \sqrt{\mu}$.

La distribución de Poisson difiere de una distribución binomial en estas formas fundamentales:

1. La distribución binomial se afecta por el tamaño de la muestra n y la probabilidad p, mientras que la distribución de Poisson sólo se afecta por la media μ.
2. En una distribución binomial, los valores posibles de la variable aleatoria x son 0, 1, . . . , n, pero los valores posibles x de una distribución de Poisson son 0, 1, 2, . . . , sin límite superior.

EJEMPLO Bombas de la Segunda Guerra Mundial Al analizar los impactos de las bombas V1 en la Segunda Guerra Mundial, el sur de Londres se subdividió en 576 regiones, cada una con un área de 0.25 km^2. En total, 535 bombas estallaron en el área combinada de 576 regiones.

a. Si se selecciona una región aleatoriamente, calcule la probabilidad de que fuese blanco de impactos exactamente en dos ocasiones.

b. Con base en la probabilidad que se calculó en el inciso *a*, ¿cuántas de las 576 regiones se espera que reciban impactos exactamente dos veces?

SOLUCIÓN

a. Aplicamos la distribución de Poisson, ya que estamos tratando con las ocurrencias de un suceso (impactos de bombas) dentro de un intervalo (una región con un área de 0.25 km^2). El número medio de impactos por región es

$$\mu = \frac{\text{número de impactos de bomba}}{\text{número de regiones}} = \frac{535}{576} = 0.929$$

continúa

Probabilidad de un suceso que nunca ha ocurrido

Algunos sucesos son posibles, pero tan poco probables que nunca ocurren. He aquí un problema de este tipo, de gran interés para los científicos políticos: estime la probabilidad de que su voto determine al ganador de una elección presidencial de Estados Unidos. Andrew Gelman, Gary King y John Boscardin escribieron en el *Journal of the American Statistical Association* (vol. 93, núm. 441) que “el valor exacto de esta probabilidad es de interés menor, pero el número tiene implicaciones importantes para la comprensión de la ubicación óptima de los recursos de campaña, para saber si los estados y los grupos de votantes reciben su parte de atención de los presidentes en prospecto y de qué manera los modelos formales de ‘elección racional’ del comportamiento del votante pueden ser capaces de explicar por qué las personas votan”. Los autores demuestran cómo se obtiene el valor de probabilidad de 1 en 10 millones, para las elecciones cerradas.

Ya que buscamos la probabilidad de exactamente dos impactos en una región, sea $x = 2$. Utilizaremos la fórmula 4-9 de la siguiente manera:

$$P(x) = \frac{\mu^x \cdot e^{-\mu}}{x!} = \frac{0.929^2 \cdot 2.71828^{-0.929}}{2!} = \frac{0.863 \cdot 0.395}{2} = 0.170$$

La probabilidad de que una región particular sea blanco de impactos exactamente dos veces es de $P(2) = 0.170$.

b. Debido a que hay una probabilidad de 0.170 de que una región reciba impactos de bombas exactamente dos veces, esperamos que entre las 576 regiones el número de las que reciben impactos exactamente dos veces sea de $576 \cdot 0.170 = 97.9$.

En el ejemplo anterior también podemos calcular las probabilidades y los valores que se esperan para 0, 1, 2, 3, 4 y 5 impactos. (Nos detenemos en $x = 5$, porque ninguna región recibió impactos de bombas más de cinco ocasiones, y las probabilidades de $x > 5$ son 0.000, cuando se redondea a tres decimales). Esas probabilidades y tales valores que se esperan se muestran en la tabla 4-5. La cuarta columna de la tabla 4-5 describe los resultados reales de la Segunda Guerra Mundial. Hubo 229 regiones sin impactos, 211 que recibieron impactos una vez, y así sucesivamente. Ahora comparemos las frecuencias *predichas* por medio de la distribución de Poisson (tercera columna) con las frecuencias *reales* (cuarta columna), para concluir que hay un acuerdo muy alto. En este caso, la distribución de Poisson hace un buen trabajo al predecir los resultados que ocurrieron en realidad. (La sección 10-2 describe un procedimiento estadístico para determinar si tales frecuencias esperadas constituyen un buen "ajuste" de las frecuencias reales. Ese procedimiento sugiere que, en tal caso, hay un buen ajuste).

Tabla 4-5 Impactos de bombas V1 en 576 regiones del sur de Londres.

Número de impactos de bombas	Probabilidad	Número que se espera de regiones	Número real de regiones
0	0.395	227.5	229
1	0.367	211.4	211
2	0.170	97.9	93
3	0.053	30.5	35
4	0.012	6.9	7
5	0.002	1.2	1

Distribución de Poisson como aproximación de la distribución binomial

En ocasiones, la distribución de Poisson se utiliza para aproximar la distribución binomial, cuando n es grande y p es pequeña. Una regla práctica es utilizar una aproximación como éstas cuando se satisfacen las siguientes dos condiciones:

1. $n \geq 100$
2. $np \leq 10$

Si se cumplen dichas condiciones y deseamos utilizar la distribución de Poisson, como aproximación de la distribución binomial, necesitamos un valor de μ; ese valor se calcula utilizando la fórmula 4-6 (que se presenta originalmente en la sección 4-4):

Fórmula 4-6 $$\mu = np$$

EJEMPLO Juego Pick 3 de Illinois En el juego Pick 3 de Illinois, usted paga 50 centavos para seleccionar una secuencia de tres dígitos, como 911. Si participa en este juego una vez al día, calcule la probabilidad de ganar exactamente una vez en 365 días.

SOLUCIÓN El intervalo de tiempo es de 365 días, así que $n = 365$. Puesto que hay un conjunto ganador de números entre los 1000 posibles (del 000 al 999), $p = 1/1000$. Se satisfacen las condiciones $n \geq 100$ y $n\,p \leq 10$, de manera que utilizaríamos la distribución de Poisson como aproximación de la distribución binomial. Primero, necesitamos el valor de μ, que se calcula de la siguiente manera:

$$\mu = np = 365 \cdot \frac{1}{1000} = 0.365$$

Al calcular el valor de μ, ahora hacemos lo mismo con $P(1)$:

$$P(1) = \frac{\mu^x \cdot e^{-\mu}}{x!} = \frac{0.365^1 \cdot 2.71828^{-0.365}}{1!} = 0.253$$

Si aplicamos la distribución de Poisson como aproximación de la distribución binomial, encontraremos que hay una probabilidad de 0.253 de ganar exactamente una vez en 365 días. Si utilizamos la distribución binomial, obtendremos una probabilidad más precisa de 0.254, de modo que observamos que la aproximación de Poisson es bastante buena aquí.

Utilizando la tecnología

STATDISK Seleccione **Analysis** de la barra del menú principal; después, **Distribuciones de probabilidad** y luego **Poisson**. Haga clic en el botón OPTIONS y proceda a introducir el valor de la media μ. Use el mouse para desplazar a la derecha o a la izquierda los distintos valores de x.

Minitab Primero introduzca el valor deseado de x en la columna C1. Ahora seleccione **Calc** de la barra del menú principal, luego **Probability Distribution** y finalmente **Poisson**. Introduzca el valor de la media μ y C1 en la columna de entrada.

Excel Haga clic en f_x en la barra del menú principal, después seleccione la categoría **Statistical**, luego **POISSON** y haga clic en **OK**. En el cuadro del diálogo introduzca los valores de x y la media, luego 0 para "Cumulative". (Introducir 1 en "Cumulative" da como resultado la probabilidad de los valores hasta el valor que se introdujo de x, inclusive.)

TI-83 Plus Presione **2nd VARS** (para obtener **DISTR**), después seleccione la opción B: **poissonpdf(**. Ahora presione **ENTER** y después proceda a introducir μ, x (incluyendo la coma). Para μ, introduzca el valor de la media; para x introduzca el número deseado de ocurrencias.

4-5 Destrezas y conceptos básicos

Uso de una distribución de Poisson para calcular la probabilidad. *En los ejercicios 1 a 4 suponga que la distribución de Poisson se aplica; después, proceda a emplear la media dada para calcular la probabilidad indicada.*

1. Si $\mu = 2$, calcule $P(3)$.

2. Si $\mu = 0.5$, calcule $P(2)$.

3. Si $\mu = 100$, calcule $P(99)$.

4. Si $\mu = 500$, calcule $P(512)$.

En los ejercicios 5 a 12 utilice la distribución de Poisson para calcular las probabilidades que se indican.

5. Disminución radioactiva Los átomos radioactivos son inestables porque tienen demasiada energía. Cuando liberan su energía sobrante, se dice que disminuyen. Al estudiar el cesio 137, se descubrió que durante el curso de la disminución, durante 365 días, 1,000,000 de átomos radioactivos se reducen a 997,287 átomos radioactivos.

a. Calcule el número medio de átomos radioactivos que se perdieron durante la disminución de un día.

b. Calcule la probabilidad de que en un día dado disminuyan 50 átomos radioactivos.

6. Nacimientos Actualmente nacen 11 bebés cada año en la villa de Westport (con una población de 760) (según datos del National Center for Health Statistics de Estados Unidos).

a. Calcule el número medio de nacimientos por día.

b. Calcule la probabilidad de que en un día dado no haya nacimientos.

c. Calcule la probabilidad de que en un día dado haya al menos un nacimiento.

d. Con base en los resultados anteriores, ¿debe estar en guardia permanente el personal médico o hay que llamarlos cuando sea necesario? ¿Significa esto que las madres de Westport podrían no recibir la atención médica inmediata que probablemente sí recibirían en una área más poblada?

7. Muertes por coces de caballos Un ejemplo clásico de la distribución de Poisson implica el número de muertes de hombres del ejército prusiano causadas por coces de caballo, entre 1875 y 1894. Se combinaron datos de 14 cadáveres durante un periodo de 20 años; los 280 años-cadáveres incluyeron un total de 196 muertes. Después de calcular el número medio de muertes por año-cadáver, determine la probabilidad de que un año-cadáver, que se selecciona aleatoriamente, tenga el siguiente número de muertes:

a. 0 **b.** 1 **c.** 2 **d.** 3 **e.** 4

Los resultados reales incluyen las siguientes frecuencias: 0 muertes (en 144 años-cadáveres); 1 muerte (en 91 años-cadáveres); 2 muertes (en 32 años-cadáveres); 3 muertes (en 11 años-cadáveres); 4 muertes (en 2 años-cadáveres). Compare los resultados reales con los que se esperan de las probabilidades de Poisson. ¿Sirve la distribución de Poisson como una buena herramienta para predecir los resultados reales?

8. Muertes por homicidio En un año hubo 116 muertes por homicidio en Richmond, Virginia (de acuerdo con “A Classroom Note on the Poisson Distribution: A Model for Homicidal deaths in Richmond, VA for 1991” en *Mathematics and Computer Education*, de Wiston A. Richards). Para un día que se seleccionó aleatoriamente, calcule la probabilidad de que el número de muertes por homicidio sea

a. 0 **b.** 1 **c.** 2 **d.** 3 **e.** 4

Compare las probabilidades calculadas con los siguientes resultados reales: 268 días (ningún homicidio); 79 días (1 homicidio); 17 días (2 homicidios); 1 día (3 homicidios); no hubo días con más de 3 homicidios.

9. Ruleta Scott apuesta el número 7 para cada uno de 200 giros de una ruleta. Como $P(7) = 1/38$, él espera ganar aproximadamente cinco veces.
   a. Calcule la probabilidad de ningún triunfo en los 200 giros.
   b. Calcule la probabilidad de al menos un triunfo en los 200 giros.
   c. Scott perderá dinero si el número de triunfos es 0, 1, 2, 3, 4 o 5. Calcule la probabilidad de que Scott pierda dinero después de 200 giros.
   d. ¿Cuál es la probabilidad de que Scott obtenga alguna ganancia después de 200 giros?

10. Terremotos Durante un periodo que comprende los últimos 100 años hubo 93 terremotos importantes en el mundo (al menos 6.0 en la escala de Richter) (según datos del *World Almanac and Book of Facts).* Suponiendo que la distribución de Poisson es un modelo adecuado, calcule el número medio de terremotos importantes por año; después, la probabilidad de que el número de terremotos en un año que se selecciona al azar sea
   a. 0 **b.** 1 **c.** 2 **d.** 3 **e.** 4 **f.** 5 **g.** 6 **h.** 7

   Los resultados reales son: 47 años (0 terremotos importantes); 31 años (1 terremoto importante); 13 años (2 terremotos importantes); 5 años (3 terremotos importantes); 2 años (4 terremotos importantes); 0 años (5 terremotos importantes); 1 año (6 terremotos importantes); 1 año (7 terremotos importantes). Después de comparar las probabilidades que se calcularon con los resultados reales, ¿es la distribución de Poisson un buen modelo?

4-5 Más allá de lo básico

11. Aproximación de Poisson a una binomial La distribución de Poisson puede emplearse para aproximar una distribución binomial si $n \geq 100$ y $np \leq 10$. Suponga una distribución binomial con $n = 100$ y $p = 0.1$. Es imposible obtener 101 éxitos con una distribución como ésta, aunque *podemos* calcular la probabilidad de $x = 101$ con la aproximación de Poisson. Calcule dicho valor. ¿Qué tanto coincide el resultado con la imposibilidad de que $x = 101$ en una distribución binomial?

12. Aproximación de Poisson a una binomial Para una distribución binomial con $n = 10$ y $p = 0.5$, no debemos usar la aproximación de Poisson, ya que las condiciones $n \geq 100$ y $np \leq 10$ no se satisfacen. Suponga que de cualquier manera empleamos la aproximación de Poisson. ¿Son aproximaciones inaceptables las probabilidades resultantes? ¿Por qué?

Resumen

El concepto de distribución de probabilidad es un elemento fundamental de la estadística. Una distribución de probabilidad describe la probabilidad de cada valor de una variable aleatoria. Este capítulo incluyó sólo distribuciones de probabilidad discreta, pero los siguientes capítulos abarcarán nociones de probabilidad continua. Se estudiaron los siguientes conceptos básicos:

- Una *variable aleatoria* posee valores que se determinaron al azar.
- Una *distribución de probabilidad* consiste en todos los valores de una variable aleatoria, junto con sus probabilidades correspondientes. Una distribución de probabilidad debe cumplir dos requisitos:

$$\Sigma P(x) = 1 \text{ y, para cada valor de } x, 0 \leq P(x) \leq 1.$$

- Se pueden explorar características importantes de una *distribución de probabilidad* construyendo un histograma de probabilidad, así como calculando su media y desviación estándar por medio de las siguientes fórmulas:

$$\mu = \Sigma[x \cdot P(x)]$$

$$\sigma = \sqrt{\Sigma[x^2 \cdot P(x)] - \mu^2}$$

- En una *distribución binomial*, hay dos categorías de resultados y un número fijo de ensayos independientes con una probabilidad constante. La probabilidad de x éxitos en n ensayos se calcula empleando la fórmula de probabilidad binomial, la tabla A-1, un programa de cómputo (como STATDISK, Minitab o Excel) o una calculadora TI-83 Plus.
- En una distribución binomial, la media y la desviación estándar se obtienen fácilmente calculando los valores de $\mu = np$ y $\sigma = \sqrt{npq}$.
- Una *distribución de probabilidad de Poisson* se aplica a ocurrencias de algún suceso durante un intervalo específico; sus probabilidades se calculan con la fórmula 4-9.
- *Resultados poco comunes:* Este capítulo puso énfasis en la importancia de interpretar los resultados a través de la distinción entre los que son comunes y los que son poco comunes. Utilizamos dos criterios diferentes. Con la regla práctica del intervalo tenemos

$$\textbf{valor máximo común} = \mu + 2\sigma$$

$$\textbf{valor mínimo común} = \mu - 2\sigma$$

También podemos determinar si los resultados son poco comunes por medio de los valores de probabilidad.

Inusualmente alto: x éxitos en n ensayos es un número *inusualmente alto* de éxitos si P (x o más) es muy pequeño (tanto como 0.05 o menos).

Inusualmente bajo: x éxitos en n ensayos es un número *inusualmente bajo* de éxitos si P (x o menos) es muy pequeño (tanto como 0.05 o menos).

Ejercicios de repaso

x	$P(x)$
0	0.08
1	0.05
2	0.10
3	0.13
4	0.15
5	0.21
6	0.09
7	0.19

1. **a.** ¿Qué es una variable aleatoria?
b. ¿Qué es una distribución de probabilidad?
c. Una gráfica del *USA Today* listó los porcentajes del número de días en una semana que los adultos estadounidenses cocinan en su casa, durante una semana promedio. Por ejemplo, el 13% de los adultos estadounidenses cocinan en su casa tres días, en una semana promedio. La tabla al margen se basa en la gráfica. ¿Describe dicha tabla una distribución de probabilidad? ¿Por qué?
d. Suponiendo que la tabla describe una distribución de probabilidad, calcule su media.
e. Suponiendo que la tabla describe una distribución de probabilidad, calcule su desviación estándar.
f. ¿Será poco común seleccionar aleatoriamente a un adulto estadounidense para averiguar que no cocina en su casa durante una semana promedio? ¿Por qué?

2. Audiencia de televidentes El programa de televisión *West Wing* tiene una audiencia de 15, es decir, mientras se está transmitiendo, el 15% de los televisores sintonizan ese programa (según datos de Nielsen Media Research). Un grupo especial consta de 20 hogares que se seleccionaron al azar (cada uno de ellos con el televisor funcionado durante la transmisión del programa *West Wing*).

continúa

a. ¿Cuál es el número que se espera de televisores que se sintonizan en el programa *West Wing*?
b. En tales grupos de 20, ¿cuál es el número medio de televisores que sintonizan el programa *West Wing*?
c. En tales grupos de 20, ¿cuál es la desviación estándar del número de televisores que sintonizan el programa *West Wing*?
d. Para tal grupo de 20, calcule la probabilidad de que exactamente cinco televisores sintonicen el programa *West Wing*?
e. Para tal grupo de 20, ¿sería poco común descubrir que ningún televisor sintoniza el programa *West Wing*? ¿Por qué?

3. Prueba de drogas a empleados De las compañías que construyen carreteras y puentes, el 80% prueba a sus empleados con respecto al abuso de sustancias (según datos de la Construction Financial Management Association). Un estudio implica la selección aleatoria de 10 compañías de este tipo.
a. Calcule la probabilidad de que 5 de las 10 compañías efectúen pruebas de abuso de sustancias.
b. Calcule la probabilidad de que al menos la mitad de las compañías hagan pruebas de abuso de sustancias.
c. Para tales grupos de 10 compañías, calcule la media y la desviación estándar del número (entre 10) que efectúa pruebas de abuso de sustancias.
d. ¿Sería poco común descubrir que seis de 10 compañías hacen pruebas de abuso de sustancias? ¿Por qué?

4. Razones de despido La incapacidad para llevarse bien con otras personas es la razón que se cita en el 17% de los despidos de trabajadores (de acuerdo con datos de Robert Half International, Inc.). Con preocupación por las condiciones de trabajo de su compañía, el gerente de personal de la Boston Finance Company planea investigar los cinco despidos que ocurrieron durante el año anterior.
a. Suponiendo que se aplica la tasa del 17%, calcule la probabilidad de que, de esos cinco empleados, el número de despidos por la incapacidad de llevarse bien con otras personas sea de al menos cuatro.
b. Si el gerente de personal realmente descubre que al menos cuatro de los despidos se deben a la incapacidad de llevarse bien con otras personas, ¿será esta compañía muy diferente de otras compañías comunes? ¿Por qué?

5. Muertes Actualmente, un promedio de siete residentes del pueblo de Westport (población 760) mueren cada año (según datos del National Center for Health Statistics de Estados Unidos).
a. Calcule el número medio de muertes por día.
b. Calcule la probabilidad de que en un día dado no haya muertes.
c. Calcule la probabilidad de que en un día dado haya una muerte.
d. Calcule la probabilidad de que en un día dado haya más de una muerte.
e. Con base en los resultados anteriores, ¿debería Westport tener un plan de contingencia para manejar más de una muerte diaria? ¿Por qué?

Ejercicios de repaso acumulativo

1. Distancias de jonrones: análisis de los últimos dígitos La tabla al margen incluye los últimos dígitos de las 73 distancias que se publicaron (en pies) de los jonrones que logró Barry Bonds en el 2001, cuando estableció el récord del mayor número de jonrones en una temporada (según datos de *USA Today*). En ocasiones es posible emplear los últimos dígitos de un conjunto de datos para determinar si éstos se midieron o simplemente se reportaron. La presencia desproporcionada de los dígitos 0 y 5, suele ser un indicador seguro de que los datos se reportaron en lugar de medirse.

x	f
0	47
1	3
2	1
3	0
4	3
5	11
6	3
7	3
8	1
9	1

continúa

a. Calcule la media y la desviación estándar de los últimos dígitos.
b. Construya la tabla de frecuencias relativas que corresponde a la tabla de frecuencias dada.
c. Construya una tabla para la distribución de probabilidad de dígitos que se seleccionaron al azar, con posibilidades iguales. Liste los valores de la variable aleatoria x (0, 1, 2, . . . , 9), junto con sus probabilidades correspondientes (0.1, 0.1, 0.1, . . . , 0.1); después, calcule la media y la desviación estándar de tal distribución de probabilidad.
d. Reconociendo que los datos muestrales se desvían naturalmente de los resultados que se esperan teóricamente, ¿habrá un acuerdo burdo de los últimos dígitos dados con la distribución que esperamos con una selección aleatoria? ¿Habrá algo en los datos muestrales (como una desproporción debida a una mayor cantidad de dígitos 0 y 5) que sugiera que los últimos dígitos dados no son aleatorios? (En el capítulo 10 presentaremos un método para responder preguntas como éstas de forma más objetiva).

2. Prueba de contaminación de automóviles La Environmental Protection Agency realizó una prueba en el tubo de escape de 116,667 automóviles para determinar cuáles generaban una gran cantidad de contaminación. Se estima que el 1% de los automóviles no pasan esa prueba.
a. Si seleccionamos al azar 20 automóviles del grupo de 116,667, ¿cuántos se esperaría que no pasaran la prueba del tubo de escape?
b. Calcule la media y la desviación estándar del número de automóviles, en grupos de 20, que no pasan la prueba del tubo de escape.
c. Calcule la probabilidad de que, en un grupo de 20 automóviles que se seleccionaron aleatoriamente, haya al menos uno que no pase la prueba del tubo de escape.
d. ¿Es poco común encontrar que, en un grupo de 20 automóviles que se seleccionaron al azar, haya tres que no pasen la prueba del tubo de escape?
e. Si se seleccionan al azar dos automóviles diferentes, calcule la probabilidad de que ambos no pasen la prueba del tubo de escape.

Actividades de cooperación en equipo

1. Actividad en clase En el capítulo 1 presentamos varios ejemplos de conjuntos de datos confusos. Suponga que deseamos identificar la distribución de probabilidad del número de hijos que nacieron de parejas que se seleccionaron aleatoriamente. Pregunte a cada estudiante de la clase el número de hermanos y hermanas que tiene; ahora, registre el número total de hijos (incluyendo al estudiante) de cada familia. Construya una tabla de frecuencias relativas con el resultado que se obtuvo (los valores de la variable aleatoria x serán 1, 2, 3, . . .). ¿Cuál sería el problema si se utilizara esta tabla de frecuencias relativas como un estimado de la distribución de probabilidad del número de hijos que nacieron de parejas que se seleccionaron al azar?

2. Actividad en clase Divídanse en equipos de tres. Seleccionen a una persona a quien probarán su percepción extrasensorial (PES), tratando de identificar correctamente un dígito (0, 1, 2, . . . , 9) que se seleccionará al azar por otro miembro del equipo. Otro participante del equipo debe registrar el dígito que se seleccionará al azar, el dígito que adivinará el sujeto, así como si la adivinación fue correcta o incorrecta. Construyan la tabla de la distribución de probabilidad para dígitos que se generan aleatoriamente, la tabla de frecuencias relativas para dígitos aleatorios seleccionados realmente y una tabla de frecuencias relativas para las adivinaciones. Después de comparar las tres tablas, ¿qué concluyen? ¿Qué proporción de las adivinaciones fue correcta? ¿Parecería que el sujeto tiene la habilidad de seleccionar el dígito correcto, de manera significativa, con mayor frecuencia de lo que se esperaría por el azar?

Proyecto tecnológico

El vuelo 2705 de American Air, que va de Nueva York a San Francisco, incluye asientos para 340 pasajeros. En promedio, el 5% de las personas con reservaciones no se presenta, por lo que American Air vende boletos por encima del cupo y acepta 350 reservaciones para los 340 asientos. Analizamos este sistema a través de una distribución binomial con $n = 350$ y $p = 0.95$ (la probabilidad de que alguien con una reservación sí se presente).

Calcule la probabilidad de que, al aceptar 350 reservaciones en un vuelo particular, haya un mayor número de pasajeros que de asientos. Es decir, calcule la probabilidad de que al menos 341 personas con reservación se presenten, suponiendo que se aceptaron 350 reservaciones. Por el valor de n, no es posible utilizar la tabla A-1; además, los cálculos con la fórmula de probabilidad binomial serían extremadamente largos y tediosos. La mejor opción es utilizar un programa de cómputo de estadística o una calculadora TI-83 Plus. Consulte la sección 4-3 para encontrar instrucciones que describen el uso de STATDISK, Minitab, Excel o una calculadora TI-83 Plus. ¿La probabilidad de sobreventa del vuelo será suficientemente pequeña de modo que no suceda con frecuencia o parece ser demasiado alta, de modo que deben hacerse cambios para disminuirla?

de los DATOS *a la* DECISIÓN

Pensamiento crítico: ¿es mejor jugar en una máquina tragamonedas o en una ruleta?

El autor compró una máquina tragamonedas Las Vegas Mills Golden Nugget con el propósito de determinar su funcionamiento. A pesar de que esta máquina se basa en un diseño de 1940, trabaja según los mismos principios que utilizan las máquina tragamonedas que hay en los casinos de Las Vegas. Dicha máquina tiene tres ruedas que giran de forma independiente y, para cada carrete, se selecciona al azar una de 20 posiciones distintas cada vez que se jala la palanca. Las tablas siguientes resumen los posibles resultados y las posiciones ganadoras. Cada juego cuesta 25 centavos.

a. Con la información de las dos tablas, llene las probabilidades en la segunda tabla. Por ejemplo, hay cuatro resultados que ganan un premio mayor, y 8000 posibles resultados diferentes, de modo que P(premio mayor) = 4/8000.

b. ¿Describe la segunda tabla ya completa una distribución de probabilidad? ¿Por qué?

c. Sea x la variable que represente la ganancia o las pérdidas netas de un solo juego de la máquina tragamonedas y calcule la media de dicha variable aleatoria. Con base en tal resultado, ¿cuál es la cantidad promedio que se gana o pierde cuando un jugador introduce 25 centavos para un juego? ¿Cuál es la recuperación promedio para cada dólar apostado?

d. Cuando se apuesta un dólar al número 7 en la ruleta, hay una probabilidad de 1/38 de ganar, en tanto un triunfo genera una ganancia neta de \$35. ¿Cuál es la recuperación promedio por cada dólar apostado?

e. Compare los resultados de los incisos *c* y *d* para determinar si es mejor jugar en una máquina tragamonedas o apostar al 7 en la ruleta. Explique.

Frecuencia de imágenes en los tres carretes

	Carrete 1	Carrete 2	Carrete 3
Pepita de oro	2	2	1
Limón	0	0	4
Campana	1	7	7
Naranja	7	2	5
Ciruela	7	2	3
Cereza	3	7	0

Resultados posibles de los tres carretes

	Ganancia neta	Probabilidad
Premio mayor (3 pepitas de oro)	36.50	
Campana-campana-campana	4.25	
Campana-campana-pepita de oro	4.25	
Ciruela-ciruela-ciruela	3.25	
Ciruela-ciruela-pepita de oro	3.25	
Naranja-naranja-naranja	2.25	
Naranja-naranja-pepita de oro	2.25	
Cereza-cereza-cualquiera	1.00	
Cereza-no cereza-cualquiera	0.25	
Pérdida: cualquier resultado que no se incluya en los nueve renglones anteriores	−0.25	

PROYECTO DE INTERNET

Distribuciones de probabilidad y simulaciones

Las distribuciones de probabilidad se utilizan para predecir el resultado de sucesos que modelan. Por ejemplo, si lanzamos una moneda balanceada, la distribución del resultado es una probabilidad de 0.5 para las caras y 0.5 para las cruces. Si lanzamos la moneda 10 veces consecutivas, esperamos cinco caras y cinco cruces. Quizá no tengamos el resultado exacto, pero a la larga, después de cientos o miles de lanzamientos, esperamos que la división entre caras y cruces sea muy cercana a "50-50". Visite el sitio de Internet de este libro de texto:

http://www.pearsoneducacion.net/triola

Localice el proyecto de Internet del capítulo 4, donde encontrará dos exploraciones. En la primera, se le pide crear una distribución de probabilidad para un experimento sencillo y utilizar esa distribución para predecir el resultado de ensayos repetidos del experimento. En la segunda exploración, analizaremos una situación más complicada: las rutas de canicas que ruedan, mientras se mueven de forma similar al *pinball*, a través de un grupo de obstáculos. En cada caso, una simulación visual dinámica le permitirá comparar los resultados predichos con un conjunto de resultados experimentales.

La estadística @ en el trabajo

"Nuestro programa es realmente un programa de educación, pero es ampliamente reconocido debido a que los resultados se han hecho públicos".

Bárbara Carvalho

Directora del Marist College Poll

Lee Miringoff

Director del Marist College Institute for Public Opinion

Barbara Carvalho y Lee Miringoff reportan los resultados de sus encuestas en muchas entrevistas para medios impresos y electrónicos, incluyendo programas de noticias de NBC, CBS, ABC, Fox y la televisión pública. Lee Miringoff aparece regularmente en el programa *Today* de la NBC.

¿A qué se dedican?

Realizamos encuestas públicas. Hacemos encuestas sobre asuntos públicos, estimaciones de aprobación de funcionarios públicos en la ciudad y el estado de Nueva York, así como a lo largo de toda la nación. No somos partidarios de realizar encuestas para partidos políticos, candidatos políticos o grupos de poder. Recibimos fondos de manera independiente del Marist College y no recibimos ingresos externos que pudiesen sugerir que hacemos investigación para algún grupo particular o sobre un tema específico.

¿Cómo seleccionan a los individuos que encuestan?

En una encuesta estatal, seleccionamos a los sujetos en proporción a los registros de votantes de los condados. Los distintos condados tienen diferentes tasas de rechazo, por lo que, si seleccionáramos personas al azar a lo largo de todo el estado, obtendríamos un modelo desigual de éste. Hacemos estratos por condado y usamos marcación de dígitos aleatoria, de modo que obtenemos números que se incluyen y no se incluyen en el directorio telefónico.

Acaba de mencionar las tasas de rechazo, ¿constituyen éstas un verdadero problema?

Uno de los aspectos que tenemos que enfrentar constantemente es el hecho de que la gente no responde las encuestas. Este fenómeno se incrementa con el tiempo y recibe mucha atención por parte de la comunidad de investigación por encuesta. Como centro de investigación, nos va bastante bien en comparación con otros. Pero cuando se hacen entrevistas cara a cara y se tienen tasas de rechazo del 25% al 50%, hay una verdadera preocupación por descubrir quién se rehúsa, por qué no quiere responder y su impacto en la representatividad de los estudios que realizamos.

¿Recomendarían ustedes un curso de estadística para estudiantes?

Totalmente. Los números no se crean todos de la misma forma. Sin importar su campo de estudio o sus intereses profesionales, es una gran ventaja poseer la habilidad para evaluar de forma crítica la información de investigaciones que se les presenten, utilizar datos para mejorar servicios o interpretar resultados para diseñar estrategias. Las encuestas, en particular, están por todas partes. Es vital que como trabajadores, gerentes y ciudadanos seamos capaces de evaluar su precisión y valor. La estadística cubre todas las disciplinas. Los estudiantes se encontrarán con ella inevitablemente en sus carreras, en algún momento.

¿Tiene alguna otra recomendación para los estudiantes?

Es importante que los estudiantes aprovechen cualquier oportunidad para desarrollar sus habilidades de comunicación y presentación. No es suficiente mejorar sus habilidades para hablar y escribir, sino que también deben incrementar su nivel de familiaridad con las nuevas tecnologías.

5

Distribuciones de probabilidad normal

¿Cómo nos adaptamos?

Una disciplina relativamente nueva es la *ergonomía*, que estudia la adaptación de las personas a su ambiente. El buen diseño ergonómico resulta en un ambiente seguro, funcional, eficiente y cómodo. Las aplicaciones de la ergonomía incluyen el diseño de tableros de automóvil, cascos para ciclistas, tapas de botellas, perillas de puertas, cubiertas para registros, teclados, centros de control de tráfico aéreo y líneas de ensamble de computadoras. Por ejemplo, en Vail, Colorado, el teleférico que lleva a los esquiadores a la cima de la montaña tiene un letrero que especifica que la capacidad máxima es de 12 personas o 2004 libras. La lectura del anuncio hace que muchos pasajeros miren a su alrededor y se pregunten si están en peligro porque hay demasiadas personas o porque hay 12 individuos (o incluso menos) que, por variación aleatoria, son exageradamente pesados. ¿Qué tan probable es que 12 personas que se eligen al azar tengan un peso total mayor de 2004 libras?

La habilidad para tolerar un largo vuelo transcontinental se afecta por el ancho del asiento que ocupemos. La mayoría de los aviones comerciales de Estados Unidos contienen asientos que miden entre 17 y 18 pulgadas de ancho, que apenas rebasan las 16 pulgadas requeridas por un pasajero promedio. Los asientos de primera clase y de la clase de negocios suelen tener anchuras entre 19 y 21 pulgadas, de modo que un espacio más grande permite un mayor grado de comodidad. Si American Airlines quiere ganar más incrementando la comodidad de sus pasajeros, ¿qué anchura deben tener sus asientos que rediseña?

Cuando se visitan construcciones que datan de hace cientos de años, muchas personas se sorprenden por el hecho de que las entradas tienen aberturas demasiado bajas para la mayoría de los adultos actuales. Cuando caminamos a través de una entrada moderna, la mayoría de nosotros cabe cómodamente por debajo del umbral, que suele medir 80 pulgadas de alto. Sin embargo, algunas personas son excepcionalmente altas y deben agacharse para evitar golpearse la cabeza. ¿Qué porcentaje de personas son demasiado altas para los estándares de diseño de las entradas actuales?

En años recientes, la Fuerza Aérea de Estados Unidos reconoció que las mujeres son muy buenos pilotos de aviones de guerra. Las cabinas de los aviones de guerra se diseñaron originalmente para hombres, de modo que se requirieron varios cambios para acomodar mejor a las mujeres pilotos. Uno de dichos cambios implicó el rediseño de los asientos de expulsión ACES-II. Puesto que se diseñaron originalmente para hombres que pesaran entre 140 y 211 libras, los asientos de expulsión implicaban un mayor riesgo de daño para cualquier mujer piloto que pesara menos de 140 libras o más de 211 libras. ¿Qué pesos deben utilizarse para el nuevo diseño de la cabina?

En este capítulo resolvemos preguntas como las anteriores.

5-1 Panorama general

En el capítulo 2, consideramos medidas importantes de conjuntos de datos, incluyendo medidas de tendencia central y de variación, así como la distribución de los datos. En el capítulo 3, estudiamos los principios básicos de probabilidad; en el capítulo 4, presentamos los siguientes conceptos:

- Una *variable aleatoria* es una variable con un valor numérico único, que se determina al azar, para cada resultado de algún procedimiento.
- Una *distribución de probabilidad* describe la probabilidad para cada valor de la variable aleatoria.
- Una variable aleatoria *discreta* tiene un número finito de valores o un número contable de valores. Es decir, el *número* de valores posibles que x puede tomar es 0 o 1, o 2, etcétera.
- Una variable aleatoria *continua* tiene un número infinito de valores, los cuales suelen asociarse con mediciones en una escala continua, sin huecos ni interrupciones.

En el capítulo 4, consideramos únicamente las distribuciones de probabilidad *discretas*, pero en este capítulo presentamos las distribuciones de probabilidad *continuas*. Aun cuando iniciamos con una distribución uniforme, la mayor parte del capítulo se enfoca en las *distribuciones normales*. Las distribuciones normales son sumamente importantes porque ocurren con gran frecuencia en las aplicaciones reales y porque juegan un papel fundamental en los métodos de estadística inferencial. Las distribuciones normales se utilizarán frecuentemente a lo largo del libro.

Definición

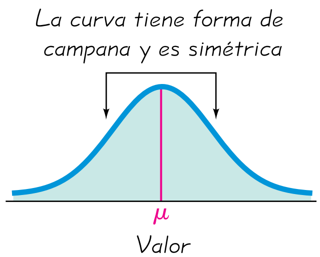
Si una variable aleatoria continua tiene una distribución con una gráfica simétrica y en forma de campana, como la de la figura 5-1, a la vez que puede ser descrita por medio de la ecuación dada como fórmula 5-1, decimos que tiene una **distribución normal**.

Fórmula 5-1

$$y = \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}}$$

La complejidad de la fórmula 5-1 provoca que muchas personas eleven las cejas mientras pronuncian la expresión "¡oh, oh!", o algo peor. Pero en realidad, te-

FIGURA 5-1 La distribución normal



nemos buenas noticias: no es necesario que utilicemos la fórmula 5-1. Sin embargo, la fórmula muestra que a cualquier distribución normal en particular la determinan dos parámetros: la media, μ, y la desviación estándar, σ. Una vez que se seleccionan valores específicos para μ y σ, se grafica la fórmula 5-1 como graficaríamos cualquier ecuación que relaciona a x con y; el resultado es una distribución de probabilidad continua con forma de campana.

5-2 Distribución normal estándar

El objetivo de este capítulo es el concepto de distribución de probabilidad normal, pero iniciamos con una *distribución uniforme*. La distribución uniforme nos permite ver algunas propiedades muy importantes que también se utilizarán con las distribuciones normales.

Distribuciones uniformes

Definición

Una variable aleatoria continua tiene una **distribución uniforme** si sus valores se dispersan uniformemente a través del rango de posibilidades. La gráfica de una distribución uniforme presenta una forma rectangular.

EJEMPLO Duración de la clase Un profesor de estadística planea sus clases con tanto cuidado que sus duraciones se distribuyen uniformemente entre 50.0 y 52.0 minutos. (Porque las clases de estadística son tan interesantes, que generalmente dan la impresión de ser más cortas). Esto es, cualquier tiempo entre 50.0 y 52.0 minutos es posible, en tanto que todos los valores posibles son igualmente probables. Si seleccionamos aleatoriamente una de las clases y permitimos que x sea la variable aleatoria que representa la duración de esa clase, entonces x tiene una distribución que puede graficarse como en la figura 5-2.

Cuando estudiamos las distribuciones de probabilidad *discretas* en la sección 4-2, identificamos dos requisitos: **1.** $\Sigma P(x) = 1$ y 2) $0 \leq P(x) \leq 1$ para todos

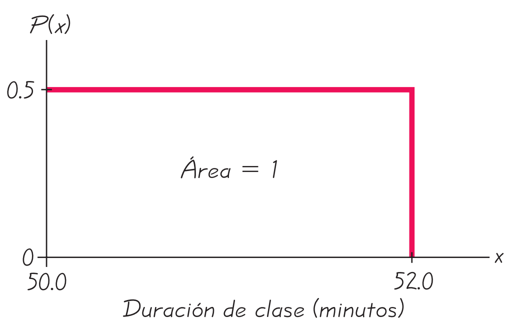


FIGURA 5-2 **Distribución uniforme de la duración de las clases**

Poblaciones cambiantes

Una de las cinco características más importantes de un conjunto de datos, que se listan en el capítulo 2, es el patrón de cambio de los datos a través del tiempo. Algunas poblaciones cambian y sus estadísticos importantes también. Los estándares de los cinturones de seguridad de los automóviles no han cambiado en 40 años, aun cuando el peso de los estadounidenses se incrementó de manera considerable desde entonces. En 1960, se consideraba que el 12.8% de los adultos estadounidenses tenían sobrepeso, en comparación con el 22.6% de 1994.

Según la National Highway Traffic Safety Administration, los cinturones de seguridad deben ajustarse a un maniquí estándar para choque (diseñado de acuerdo con los datos de 1960), que se colocó en la posición más adelante posible con 4 pulgadas de sobra. En teoría, el cinturón de seguridad tiene que ajustarse al 95% de los hombres y al 99% de las mujeres, pero tales porcentajes son ahora más bajos debido al incremento en el peso que tuvo lugar durante la última mitad del siglo. Algunas compañías proporcionan extensiones para cinturones de seguridad, pero otras no.

Muestreo rechazado para el censo

Se estima que, en el censo de 2000 en Estados Unidos, 7 millones de individuos no se contaron, mientras que otros 4 millones se consideraron dos veces. Dichos errores pueden corregirse al aplicar métodos de estadística que se conocen, aunque ello implica un asunto político. Los conteos poblacionales afectan el número de asientos en la Cámara de Representantes, de modo que los republicanos se oponen al muestreo, ya que las regiones que no se contaron completamente tienden a ser principalmente demócratas, mientras que las regiones que se cuentan en exceso tienen mayorías republicanas. Los demócratas están a favor del uso de métodos de muestreo. Algunas personas argumentan que la Constitución de Estados Unidos especifica que el censo debe ser un "conteo real" (un conteo por cabeza), que no permite métodos de muestreo; la Suprema Corte apoya esta posición. Los métodos estadísticos resultarían útiles para mejorar de manera sustancial los resultados del censo, de forma que los ciudadanos disfruten de una distribución más equitativa de la ayuda federal, junto con una representación más equitativa en el Congreso.

los valores de x. También en la sección 4-2 establecimos que la gráfica de una distribución de probabilidad discreta se denomina *histograma de probabilidad*. La gráfica de una distribución de probabilidad continua, como la que se incluye en la figura 5-2, se llama *curva de densidad*; debe satisfacer dos propiedades similares a los requisitos de las distribuciones de probabilidad discretas, tal como se plantea en la siguiente definición.

Definición

Curva de densidad (o **función de densidad de probabilidad**): gráfica de una distribución de probabilidad continua. Debe satisfacer las siguientes propiedades:

1. El área total bajo la curva debe ser igual a 1.
2. Cada punto de la curva debe tener una altura vertical igual o mayor que 0. (Es decir, la curva no puede estar por debajo del eje x).

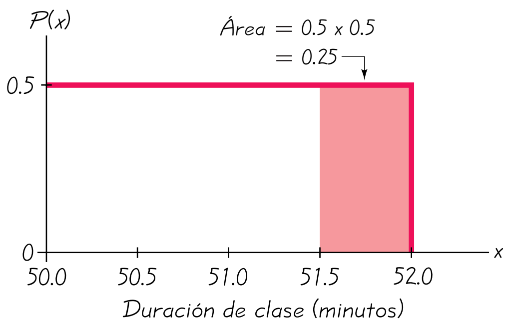
Si establecemos que la altura del rectángulo de la figura 5-2 es 0.5, obligamos a que el área circunscrita sea $2 \times 0.5 = 1$, como se requiere. (En general, el área del rectángulo se convierte en 1, cuando igualamos su altura al valor de $1/\text{rango}$). Esta propiedad (área $= 1$) facilita la solución de problemas de probabilidad, de modo que la siguiente afirmación es importante:

Ya que el área total debajo de la curva de densidad es igual a 1, hay una correspondencia entre *área* y *probabilidad.*

EJEMPLO Duración de la clase Kim, que tiene el hábito de vivir siempre de prisa, se comprometió a acudir a una entrevista de trabajo, inmediatamente después de su clase de estadística. Si la clase dura más de 51.5 minutos, llegará tarde a la entrevista. Dada la distribución uniforme de la figura 5-2, calcule la probabilidad de que una clase que se selecciona aleatoriamente dure más de 51.5 minutos.

SOLUCIÓN Observe la figura 5-3, donde la región sombreada representa duraciones mayores de 51.5 minutos. Puesto que el área total bajo la curva de

FIGURA 5-3 Uso del área para el cálculo de la probabilidad



densidad es igual a 1, hay una correspondencia entre área y probabilidad. Por lo tanto, es posible calcular la probabilidad que se desea utilizando áreas de la siguiente manera:

$$
\begin{aligned}
P\text{ (clase mayor de 51.5 minutos)} &= \text{área de región sombreada de la figura 5-3} \\
&= 0.5 \times 0.5 \\
&= 0.25
\end{aligned}
$$

INTERPRETACIÓN La probabilidad de seleccionar aleatoriamente una clase que dure más de 51.5 minutos es de 0.25. Ya que esa probabilidad es demasiado alta, Kim tiene que considerar hacer un plan de contingencia que le permita llegar a su entrevista de trabajo a tiempo. Nadie debe llegar tarde a una entrevista de trabajo.

Distribución normal estándar

La curva de densidad de una distribución uniforme es una línea horizontal, de forma que es sencillo calcular el área de cualquier región rectangular multiplicando anchura por altura. La curva de densidad de una distribución normal tiene una forma de campana más complicada, como se ve en la figura 5-1, por lo que es más difícil calcular áreas, pero el principio básico es el mismo: *existe una correspondencia entre área y probabilidad*.

Así como hay muchas distribuciones uniformes diferentes (con distintos rangos de valores), también existen muchas distribuciones normales diferentes, las cuales dependen de dos parámetros: la media poblacional, μ, y la desviación estándar poblacional, σ. (Recuerde que en el capítulo 1 vimos que un *parámetro* es una medida numérica que describe alguna característica de una *población*). La figura 5-4 incluye curvas de densidad de estaturas de hombres y mujeres adultos. Como los hombres tienen una estatura media mayor, la cima de la curva de densidad de los hombres se ubica hacia la derecha. Puesto que las estaturas de los hombres tienen una desviación estándar ligeramente mayor, su curva de densidad es un poco más ancha. La figura 5-4 presenta dos posibles distribuciones normales diferentes. Hay una infinidad de posibilidades, pero una es de especial interés.

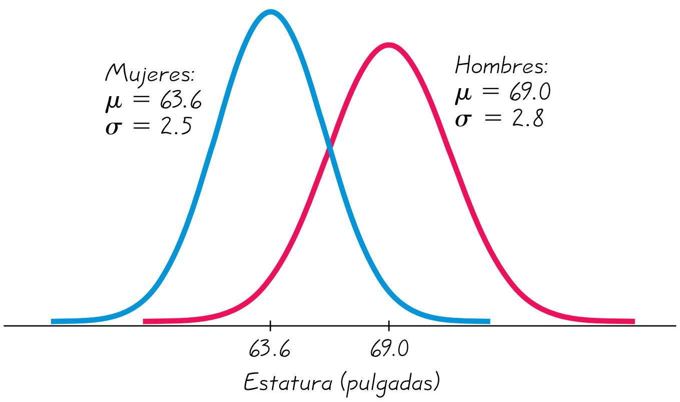


FIGURA 5-4 **Estaturas de hombres y mujeres adultos**

Confiabilidad y validez

La confiabilidad de los datos se refiere a la consistencia con que se presentan los resultados, mientras que la validez de los datos se refiere a lo bien que los datos miden lo que se supone deben medir. La confiabilidad de una prueba de CI puede juzgarse comparando las puntuaciones de la prueba en una aplicación, con las puntuaciones de la misma prueba que se aplica en otro momento. Para probar la validez de una prueba de CI, habrá que comparar las puntuaciones de la prueba con algún otro indicador de inteligencia, como el desempeño académico. Muchos críticos afirman que las pruebas de CI son confiables, pero no válidas; ofrecen resultados consistentes, aunque no miden realmente la inteligencia.

Definición

Distribución normal estándar: distribución normal de probabilidad con una media de 0 y una desviación estándar de 1, en tanto el área total debajo de su curva de densidad es igual a 1. (Véase la figura 5-5).

Suponga que nos contrataron para realizar cálculos con el uso de la fórmula 5-1. Rápidamente veríamos que los valores más fáciles para μ y para σ son $\mu = 0$ y $\sigma = 1$. Al permitir que $\mu = 0$ y $\sigma = 1$, los matemáticos calculan muchas áreas diferentes bajo la curva. Como se aprecia en la figura 5-5, el área bajo la curva es 1; lo anterior nos permite establecer la correspondencia entre área y probabilidad, tal como hicimos en el ejemplo anterior con la distribución uniforme.

Cálculo de probabilidades con puntuaciones *z*

Si empleamos la tabla A-2 (en el Apéndice A y en la tarjeta con *fórmulas y tablas*) TI-83 Plus o programas de cómputo como el STATDISK, Minitab o Excel. Las características más importantes de los distintos métodos se resumen en la tabla 5-1. No es necesario conocer los cinco métodos, sólo necesita aprender el método que utilizará para la clase y los exámenes.

Puesto que los siguientes ejemplos y ejercicios se basan en la tabla A-2, es esencial comprender los siguientes puntos:

1. La tabla A-2 se diseñó únicamente para la distribución normal *estándar*, que tiene una media de 0 y una desviación estándar de 1.
2. La tabla A-2 abarca dos páginas, una para las puntuaciones *z negativas* y otra para las puntuaciones *z positivas*.
3. Cada valor en la tabla es una *área acumulativa desde la izquierda* hasta una frontera vertical, por sobre una puntuación *z* específica.
4. Cuando construya una gráfica, evite la confusión entre puntuaciones *z* y las áreas.

   Puntuación *z*: ***Distancia* a lo largo de la escala horizontal de la distribución normal estándar; remítase a la columna de la extrema izquierda y al renglón superior de la tabla A-2.**

   Área: ***Región* bajo la curva; remítase a los valores de la tabla A-2.**

5. La parte de la puntuación *z* que denota centésimas, se encuentra en el renglón superior de la tabla A-2.

El siguiente ejemplo requiere que calculemos la probabilidad que se asocia con un valor menor que 1.58. Comience con la puntuación *z* de 1.58, localizando 1.5 en la columna izquierda; después, calcule el valor en el renglón adjunto de probabili-

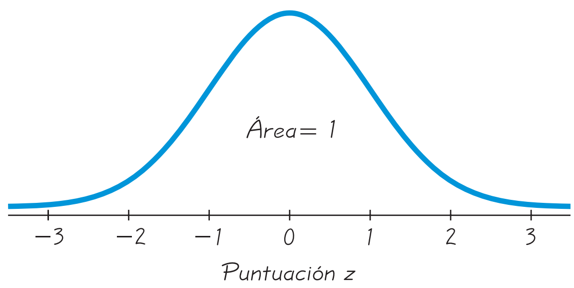
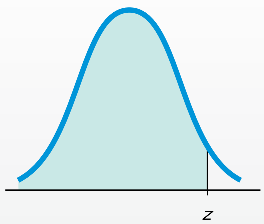
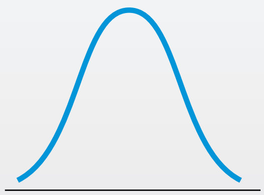
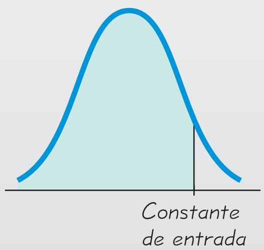
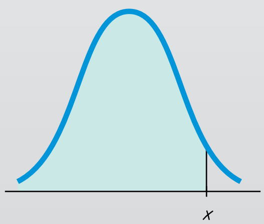
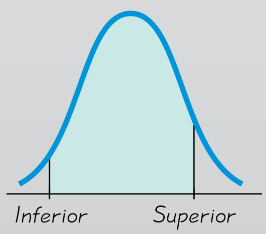


FIGURA 5-5 **Distribución normal estándar: $\mu = 0$ y $\sigma = 1$**

Tabla 5-1 Métodos para el cálculo de las áreas de la distribución normal

Tabla A-2 Da el área acumulativa de la izquierda hasta una línea vertical por encima de un valor específico de *z*.		El procedimiento para el uso de la tabla A-2 se describe en el texto.
STATDISK Da pocas áreas, incluyendo el área acumulativa de la izquierda y el área acumulativa de la derecha.		Seleccione **Analysis, Probability Distributions, Normal Distribution.** Deslice el mouse hacia la derecha y la izquierda.
Minitab Da el área acumulativa de la izquierda hasta una línea vertical por arriba de un valor específico.		Seleccione **Calc, Probability Distributions, Normal.** En el cuadro de diálogo, seleccione **Cumulative Probability, Input Constant.**
Excel Da el área acumulativa de la izquierda hasta una línea vertical por arriba de un valor específico.		Seleccione **fx, Statistical, NORMDIST.** En el cuadro de diálogo, introduzca el valor y la media, la desviación estándar y "true".
TI-83 Plus Da el área con límites izquierdo y derecho, por medio de líneas verticales, sobre cualesquiera valores específicos.		Presione `2nd` `VARS` **[2: normal cdf(]**; después, introduzca *z* separadas por una coma, como en (puntuación *z* izquierda, puntuación *z* derecha).

z	....	0.08
.		.
.		.
.		.
1.5	....	0.9429

dad que está directamente debajo de 0.08, como se observa en esta porción de la tabla A-2.

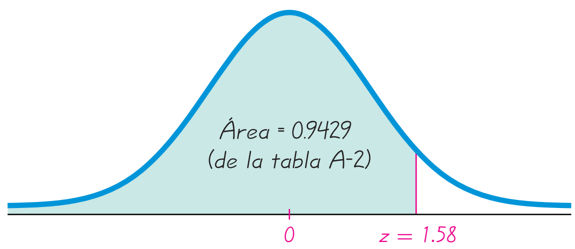
El valor del área (o probabilidad) de 0.9429 indica que hay una probabilidad de 0.9429 de seleccionar aleatoriamente una puntuación z menor que 1.58. (En las siguientes secciones, consideraremos casos en los cuales la media no es 0 ni la desviación estándar es 1.)

EJEMPLO Termómetros científicos La Precision Scientific Instrument Company fabrica termómetros que, se supone, deben dar lecturas de 0°C al punto de congelación del agua. Las pruebas de una muestra grande de tales instrumentos reveló que, en el punto de congelación del agua, algunos termómetros daban lecturas por debajo de 0° (que se denotan con números negativos) y otros por encima de 0° (que se denotan con números positivos). Suponga que la lectura media es 0°C y la desviación estándar de las lecturas es 1.00°C. También, que las lecturas se distribuyen de manera normal. Si se elige al azar un termómetro, calcule la probabilidad de que, al punto de congelación del agua, la lectura sea menor que 1.58°.

SOLUCIÓN La distribución de probabilidad de las lecturas es una distribución normal estándar, ya que las lecturas se distribuyen de forma normal, con $\mu = 0$ y $\sigma = 1$. Necesitamos encontrar el área que está debajo de $z = 1.58$, en la figura 5-6. El *área* por debajo de $z = 1.58$ es igual a la *probabilidad* de seleccionar al azar un termómetro con una lectura menor que 1.58°. En la tabla A-2 encontramos que dicha área es de 0.9429.

INTERPRETACIÓN La *probabilidad* de seleccionar aleatoriamente un termómetro con una lectura menor que 1.58°, en el punto de congelación del agua, es igual al *área* de 0.9429, que aparece como la región que se sombreó en la figura 5-6. Otra forma de interpretar el resultado es concluyendo que el 94.29% de los termómetros tendrán lecturas por debajo de 1.58°.

FIGURA 5-6 Cálculo del área por debajo de $z = 1.58$



EJEMPLO Termómetros científicos Utilice los termómetros del ejemplo anterior y calcule la probabilidad de seleccionar aleatoriamente un termómetro con una lectura, en el punto de congelación del agua, por arriba de −1.23°.

SOLUCIÓN Nuevamente, calculamos la *probabilidad* que se desea encontrando un *área* correspondiente. Buscamos el área de la región que se sombreó en la figura 5-7, pero la tabla A-2 se diseñó para aplicarse sólo en áreas acumulativas desde la *izquierda*. Si nos remitimos a la tabla A-2, en la página con puntuaciones z negativas, encontramos que el área acumulativa de la izquierda hasta $z = -1.23$ es 0.1093, como se observa. Sabiendo que el área total bajo la

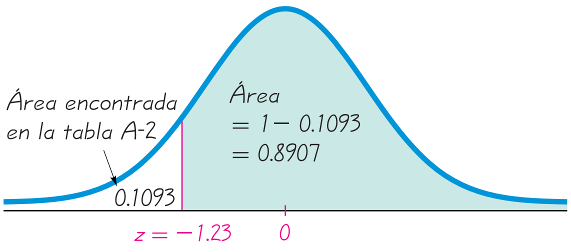


FIGURA 5-7 cálculo del área por encima de $z = -1.23$

curva es 1, calcularemos el área que se sombreó si restamos 0.1093 de 1. El resultado es 0.8907. Aun cuando la tabla A-2 se diseñó únicamente para áreas acumulativas de la izquierda, podemos utilizarla para calcular áreas acumulativas de la derecha, tal como se muestra en la figura 5-7.

INTERPRETACIÓN Por la correspondencia entre probabilidad y área, concluimos que la *probabilidad* de seleccionar aleatoriamente un termómetro con una lectura por arriba de −1.23°, en el punto de congelación del agua, es de 0.8907 (correspondiente al *área* que está por arriba de $z = -1.23$). En otras palabras, el 89.07% de los termómetros tienen lecturas por encima de −1.23°.

El ejemplo anterior ilustra una de las formas en que es posible utilizar la tabla A-2 para calcular, de manera indirecta, una área acumulativa de la derecha. El siguiente ejemplo ilustra otra manera para calcular el área con el uso de la tabla A-2.

EJEMPLO Termómetros científicos Una vez más, haga una selección aleatoria de la misma muestra de termómetros y calcule la probabilidad de que el termómetro que se eligió tenga lecturas, en el punto de congelación del agua, entre −2.00° y 1.50°.

SOLUCIÓN Nuevamente tratamos con valores que se distribuyen de manera normal, con una media de 0° y una desviación estándar de 1°. La probabilidad de seleccionar un termómetro con lecturas entre −2.00° y 1.50° corresponde al área que se sombreó en la figura 5-8. La tabla A-2 no puede utilizarse para calcular el área de forma directa, pero sí para encontrar que $z = -2.00$ corresponde al área de 0.0228, y que $z = 1.50$ corresponde al área de 0.9332, como se indica en la figura. Remítase a la figura 5-8 y observe que el área que se sombreó corresponde a la diferencia entre 0.9332 y 0.0228. El área que se sombreó es, por lo tanto, $0.9332 - 0.0228 = 0.9104$.

continúa

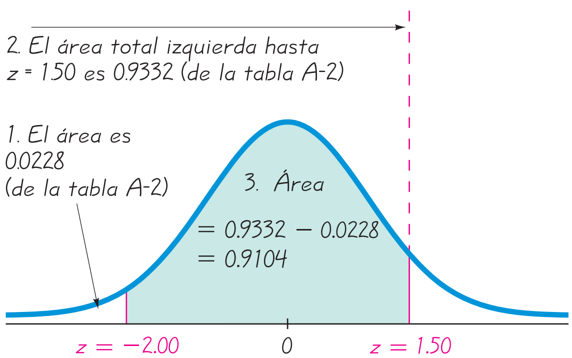


FIGURA 5-8 Cálculo del área entre dos valores

INTERPRETACIÓN Con el uso de la correspondencia entre probabilidad y área, concluimos que hay una probabilidad de 0.9104 de seleccionar aleatoriamente uno de los termómetros con una lectura de entre −2.00° y 1.50°, en el punto de congelación del agua. Otra forma de interpretar este resultado es afirmar que si se seleccionan muchos termómetros, y se prueban en el punto de congelación del agua, entonces 0.9104 (o el 91.04%) de ellos tendrán lecturas entre −2.00° y 1.50°.

El ejemplo anterior puede generalizarse como una regla que establece que el área correspondiente a la región que se localiza entre dos puntuaciones z específicas se obtiene calculando la diferencia entre las dos áreas que se localizan en la tabla A-2. Observe la figura 5-9, que muestra que la región B que se sombreó se obtiene calculando la *diferencia* entre dos áreas de la tabla A-2: las áreas A y B combinadas (que en la tabla A-2 aparecen como las áreas correspondientes a z_{Derecha}) y el área A (que en la tabla A-2 aparece como el área correspondiente a $z_{\text{Izquierda}}$). *Sugerencia*: No trate de memorizar una regla o una fórmula para este caso, ya que es infinitamente mejor *comprender* el procedimiento. Entienda, mejor, cómo funciona la tabla A-2; después, dibuje una gráfica, sombree el área que se desea y piense en una forma para calcular el área, considerando que la tabla A-2 proporciona sólo áreas acumulativas desde la izquierda.

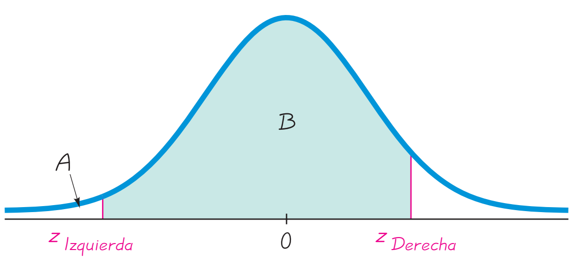
El ejemplo anterior concluyó con la afirmación de que la probabilidad de una lectura entre −2.00° y 1.50° es de 0.9104. Probabilidades como ésta, también pueden expresarse con la siguiente notación:

Notación

$P(a < z < b)$	denota la probabilidad de que la puntuación z esté entre a y b.
$P(z > a)$	denota la probabilidad de que la puntuación z sea mayor que a.
$P(z < a)$	denota la probabilidad de que la puntuación z sea menor que a

Con el uso de esa notación expresaremos el resultado del último ejemplo de la siguiente manera: $P(-2.00 < z < 1.50) = 0.9104$ que, en símbolos, establece que la probabilidad de que una puntuación z caiga entre −2.00 y 1.50 es de 0.9104. Con una distribución de probabilidad continua, tal como la distribución normal, la probabilidad de obtener cualquier valor *exacto* es de 0. Es decir, $P(z = a) = 0$.

FIGURA 5-9 Cálculo del área entre dos puntuaciones z



Área sombreada B = (áreas A y B combinadas) − (área A) = (área de la tabla A-2, usando z_{Derecha}) − (área de la tabla A-2, usando $z_{\text{Izquierda}}$)

Por ejemplo, hay una probabilidad 0 de seleccionar aleatoriamente a alguien y obtener una persona cuya estatura sea con exactitud de 68.12345678 pulgadas. En la distribución normal, cualquier punto único sobre la escala horizontal se representa, no por una región bajo la curva, sino por una línea vertical por encima del punto. Para $P(z = 1.50)$, tenemos una línea vertical que está por arriba de $z = 1.50$, pero la línea vertical, por sí misma, no contiene un área, de modo que $P(z = 1.50) = 0$. Para cualquier variable aleatoria continua la probabilidad de un valor exacto es 0; además, se infiere que $P(a \leq z \leq b) = P(a < z < b)$. También se deduce que la probabilidad de obtener una puntuación z de *a lo sumo b*, es igual a la probabilidad de obtener una puntuación z *menor que b*. Es importante interpretar correctamente frases clave como *a lo sumo, al menos, mayor que, no mayor que*, etcétera.

Cálculo de puntuaciones *z* de áreas conocidas

Hasta ahora, todos los ejemplos de esta sección que implican la distribución normal estándar siguen el mismo formato: dadas puntuaciones z, calculamos áreas bajo la curva; dichas áreas corresponden a probabilidades. En muchos otros casos, realizamos el proceso contrario, porque ya conocemos el área (o probabilidad), pero necesitamos calcular la puntuación z correspondiente. En estos casos, es muy importante evitar una confusión entre las puntuaciones z y las áreas. Recuerde, las puntuaciones z son *distancias* a lo largo de la escala horizontal, que se representa con los números de la tabla A-2, que se encuentran en la columna de la extrema izquierda y en el cruce del renglón superior. Las áreas (o probabilidades), regiones bajo la curva, se representan con los valores en el cuerpo de la tabla A-2. Asimismo, las puntuaciones z que se ubican en la mitad izquierda de la curva siempre son negativas. Si ya conocemos una probabilidad y deseamos determinar la puntuación z correspondiente, la calculamos de la siguiente forma.

Procedimiento para el cálculo de una puntuación *z* a partir de un área conocida

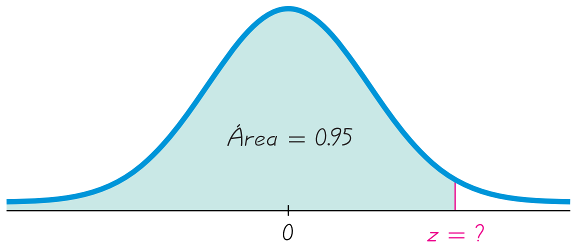
1. Dibuje una curva en forma de campana e identifique la región bajo la curva correspondiente a la probabilidad dada. Si no se trata de una región acumulativa de la izquierda, en su lugar trabaje con una región acumulativa que se conoce de la izquierda.
2. Usando el área acumulativa de la izquierda, localice la probabilidad más cercana en el *cuerpo* de la tabla A-2 e identifique la puntuación z correspondiente.

EJEMPLO Termómetros científicos Use los mismos termómetros anteriores, con lecturas de temperatura al punto de congelación del agua que se distribuyen normalmente, con una media de 0°C y una desviación estándar de 1°C. Calcule la temperatura correspondiente a P_{95}, el percentil 95. Es decir, determine la temperatura que separa el 95% inferior del 5% superior. Observe la figura 5.10.

SOLUCIÓN La figura 5-10 incluye la puntuación z correspondiente al percentil 95, con el 95% del área (o 0.95) por debajo de ella. Importante: Cuando se remita a la tabla A-2, recuerde que el cuerpo de la tabla incluye las *áreas acumulativas de la izquierda*. Al remitirnos a la tabla A-2, buscamos el área de

continúa

FIGURA 5-10 Cálculo del percentil 95



0.95 *en el cuerpo* de la tabla y después calculamos la puntuación z correspondiente. En la tabla encontramos las áreas de 0.9495 y 0.9505, donde hay un asterisco con una nota especial que indica que 0.9500 corresponde a una puntuación z de 1.645. Entonces concluimos que la puntuación z, en la figura 5-10, es 1.645, por lo que el percentil 95 es la lectura de la temperatura de 1.645°C.

INTERPRETACIÓN Al realizar pruebas a la temperatura de congelación, el 95% de las lecturas serán menores o iguales que 1.645°C, en tanto que el 5% de ellas será mayor o igual que 1.645°C.

Puntuación z	Área acumulativa de la izquierda
1.645	0.9500
−1.645	0.0500
2.575	0.9950
−2.575	0.0050

Note que en la solución anterior, la tabla A-2 indicó una puntuación z de 1.645, que está a la mitad de 1.64 y 1.65. Con la tabla A-2, generalmente evitaremos la interpolación si seleccionamos sencillamente el valor más cercano. Hay casos especiales, que se listan en la tabla adjunta, los cuales son importantes porque se utilizan con frecuencia en una amplia variedad de aplicaciones. (El valor de $z = 2.576$ da un área ligeramente más cercana a la de 0.9950, pero $z = 2.575$ tiene la ventaja de ser el valor intermedio entre $z = 2.57$ y $z = 2.58$). Con la excepción de estos casos especiales, es posible seleccionar el valor más cercano en la tabla. (Si un valor que se desea se encuentra entre dos valores de la tabla, seleccione el valor más grande). Además, para las puntuaciones z por arriba de 3.49, utilizaremos 0.9999 como aproximación del área acumulativa de la izquierda; para puntuaciones z por debajo de −3.49, usaremos 0.0001 como aproximación del área acumulativa de la izquierda.

EJEMPLO Termómetros científicos Utilice los mismos termómetros y calcule las temperaturas que separan el 2.5% inferior y el 2.5% superior.

SOLUCIÓN Remítase a la figura 5-11, que presenta las puntuaciones z que se requieren. Para encontrar la puntuación z que se localiza a la izquierda, remítase a la tabla A-2 y busque un área de 0.025, en el *cuerpo de la tabla*. El resultado es $z = -1.96$. Para encontrar la puntuación z que se localiza a la derecha, remítase al *cuerpo de la tabla* y busque un área de 0.975. (Recuerde que la tabla A-2 siempre da áreas acumulativas de la *izquierda*). El resultado es $z = 1.96$. Los valores de $z = -1.96$ y $z = 1.96$ separan el 2.5% inferior y el 2.5% superior, como muestra la figura 5-11.

INTERPRETACIÓN Al realizar pruebas a la temperatura de congelación, el 2.5% de las lecturas de los termómetros serán iguales o menores que −1.96°, en tanto que el 2.5% de las lecturas serán iguales o mayores que 1.96°. Otra interpretación es que, al punto de congelación del agua, el 95% de todas las lecturas de los termómetros se ubicarán entre −1.96° y 1.96°.

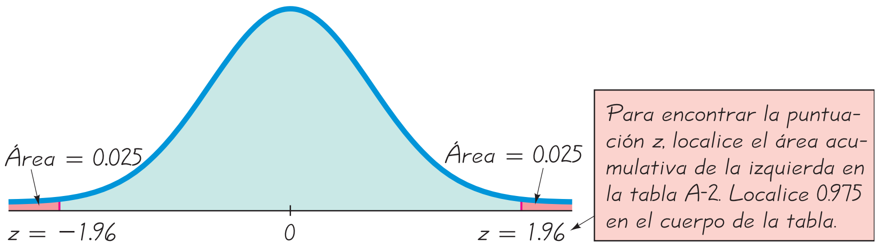


FIGURA 5-11 Cálculo de puntuaciones z

Los ejemplos de esta sección se elaboraron de forma que la media de 0 y la desviación estándar de 1 coincidieran exactamente con los parámetros de la distribución normal estándar. En realidad, es raro encontrar parámetros tan convenientes, ya que las distribuciones normales clásicas incluyen medias distintas de 0 y desviaciones estándar distintas de 1. En la siguiente sección, introducimos métodos para trabajar con este tipo de distribuciones normales, que son más realistas.

Utilizando la tecnología

STATDISK Seleccione **Analysis, Probability Distributions, Normal Distribution**; luego, proceda a deslizar el mouse hacia la derecha o la izquierda, hasta encontrar el valor que se desea. Quizá logre mayor precisión si emplea el mouse para arrastrar parte de la curva, de manera que sea posible amplificarla.

Minitab

- Para encontrar el área acumulativa que está a la izquierda de una puntuación z (como en la tabla A-2), seleccione **Calc, Probability Distributions, Normal, Cumulative probabilities;** después, introduzca la media de 0 y la desviación estándar de 1, haga clic en el botón de **Input Constant** e inserte la puntuación z.
- Para encontrar la puntuación z correspondiente a una probabilidad que se conoce, seleccione **Calc, Probability Distributions, Normal**; después, elija **Inverse cumulative probabilities** y la opción **Input constant.** Para la constante de entrada, introduzca el área total que se encuentra a la izquierda del valor dado.

Excel

- Para encontrar el área acumulativa a la izquierda de una puntuación z (como en la tabla A-2), haga clic en ***fx****;* después, seleccione **Statistical, NORMSDIST** e introduzca la puntuación z.
- Para encontrar la puntuación z correspondiente a una probabilidad que se conoce, seleccione ***fx*****, Statistical, NORMSINV** e introduzca el área total que se encuentra a la izquierda del valor dado.

TI-83 Plus

- Para calcular el área entre dos puntuaciones z, presione **2nd VARS, 2**(para normalcdf); después, proceda a introducir las dos puntuaciones z, que se separaron con una coma, como en (puntuación z izquierda, puntuación z derecha).
- Para encontrar una puntuación z correspondiente a una probabilidad que se conoce, presione **2nd VARS, 3** (para invNorm), y proceda a introducir el área total a la izquierda del valor, la media y la desviación estándar con el formato (área total izquierda, media, desviación estándar), incluyendo las comas.

5-2 Destrezas y conceptos básicos

Uso de una distribución uniforme continua*. En los ejercicios 1 a 4, remítase a la distribución uniforme de la figura 5-2; suponga que se selecciona una clase con duración entre 50.0 y 52.0 minutos, calcule la probabilidad de seleccionar el tiempo dado.*

1. Menor que 50.3 min.

2. Mayor que 51.0 min.

3. Entre 50.5 minutos y 50.8 minutos.

4. Entre 50.5 minutos y 51.8 minutos.

***Uso de una distribución uniforme continua**. En los ejercicios 5 a 8, suponga que los voltajes en un circuito varían entre 6 y 12 volts, así como que los voltajes se distribuyen de forma equitativa en el rango de posibilidades, de modo que existe una distribución uniforme. Calcule la probabilidad del rango dado de niveles de voltaje.*

5. Mayor que 10 volts.

6. Menor que 11 volts.

7. Entre siete y 10 volts.

8. Entre 6.5 y ocho volts.

***Uso de la distribución normal estándar.** En los ejercicios 9 a 28, suponga que las lecturas de los termómetros se distribuyen normalmente, con una media de 0° y una desviación estándar de 1.00°C. Se selecciona aleatoriamente un termómetro y se prueba. En cada caso, dibuje un bosquejo y calcule la probabilidad de cada lectura en grados.*

9. Menor que -0.25.

10. Menor que -2.75.

11. Menor que 0.25.

12. Menor que 2.75.

13. Mayor que 2.33.

14. Mayor que 1.96.

15. Mayor que -2.33.

16. Mayor que -1.96.

17. Entre 0.50 y 1.50.

18. Entre 1.50 y 2.50.

19. Entre -2.00 y -1.00.

20. Entre 2.00 y 2.34.

21. Entre -2.67 y 1.28.

22. Entre -1.18 y 2.15

23. Entre -0.52 y 3.75.

24. Entre -3.88 y 1.07.

25. Mayor que 3.57.

26. Menor que -3.61.

27. Mayor que 0.

28. Menor que 0.

***Bases de la regla empírica.** En los ejercicios 29 a 32, calcule el área bajo la curva que se indica de la distribución normal estándar; después, conviértala en porcentaje y llene el espacio en blanco. Los resultados conforman la base de la regla empírica que se explicó en la sección 2-5.*

29. Aproximadamente el ____ % del área está entre $z = -1$ y $z = 1$ (o dentro de una desviación estándar a partir de la media).

30. Aproximadamente el ____ % del área está entre $z = -2$ y $z = 2$ (o dentro de dos desviaciones estándar a partir de la media).

31. Aproximadamente el ____ % del área está entre $z = -3$ y $z = 3$ (o dentro de tres desviaciones estándar a partir de la media).

32. Aproximadamente el ____ % del área está entre $z = -3.5$ y $z = 3.5$ (o dentro de 3.5 desviaciones estándar a partir de la media).

***Cálculo de probabilidad.** En los ejercicios 33 a 36, suponga que las lecturas de los termómetros se distribuyen normalmente, con una media de 0° y una desviación estándar de 1.00°. Calcule la probabilidad que se indica, donde z es la lectura en grados.*

33. $P(-1.96 < z < 1.96)$

34. $P(z < 1.645)$

35. $P(z > -2.575)$

36. $P(1.96 < z < 2.33)$

Cálculo de valores de temperatura. *En los ejercicios 37 a 40, suponga que las lecturas de los termómetros se distribuyen normalmente, con una media de 0° y una desviación estándar de 1.00°C. Se selecciona aleatoriamente un termómetro y se prueba. En cada caso, dibuje un bosquejo y calcule la lectura de la temperatura correspondiente a la información dada.*

37. Calcule P_{90}, el percentil 90o. Ésta es la lectura de temperatura que separa el 90% inferior del 10% superior.

38. Calcule P_{20}, el percentil 20o.

39. Si se rechaza el 5% de los termómetros porque tienen lecturas muy bajas, pero el resto de los termómetros son aceptables, calcule la lectura que separa a los termómetros que se rechazaron de los otros.

40. Si se rechaza el 3.0% de los termómetros, porque tienen lecturas muy altas, y otro 3.0% se rechaza por registrar lecturas muy bajas, calcule las dos lecturas de los valores que separan a los termómetros que se rechazaron de los otros.

5-2 Más allá de lo básico

41. Para una distribución normal estándar, calcule el porcentaje de datos que están:
a. Dentro de 1 desviación estándar a partir de la media.
b. Dentro de 1.96 desviaciones estándar a partir de la media.
c. Entre $\mu - 3\sigma$ y $\mu + 3\sigma$.
d. Entre 1 desviación estándar por debajo de la media y 2 desviaciones estándar por encima de la media.
e. A más de 2 desviaciones estándar a partir de la media.

42. Si una distribución uniforme continua tiene parámetros de $\mu = 0$ y $\sigma = 1$, entonces el mínimo es $-\sqrt{3}$ y el máximo es $\sqrt{3}$.
a. Para esta distribución calcule $P(-1 < x < 1)$.
b. Calcule $P(-1 < x < 1)$ si considera de manera incorrecta que la distribución es normal en lugar de uniforme.
c. Compare los resultados de los incisos *a* y *b*. ¿Afecta mucho la distribución a los resultados?

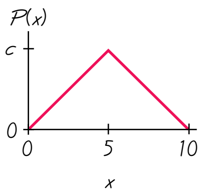
43. Suponga que puntuaciones *z* se distribuyen normalmente, con una media de 0 y una desviación estándar de 1.
a. Si $P(0 < z < a) = 0.3907$, calcule a.
b. Si $P(-b < z < b) = 0.8664$, calcule b.
c. Si $P(z > c) = 0.0643$, calcule c.
d. Si $P(z > d) = 0.9922$, calcule d.
e. Si $P(z < e) = 0.4500$, calcule e.

44. En una distribución uniforme continua,

$$\mu = \frac{\text{mínimo} + \text{máximo}}{2} \qquad \text{y} \qquad \sigma = \frac{\text{rango}}{\sqrt{12}}$$

Calcule la media y la desviación estándar de la distribución uniforme que se representa en la figura 5-2.

45. Realice el dibujo de una gráfica que represente una distribución acumulativa de *a* una distribución uniforme y *b* una distribución normal.



46. Remítase a la gráfica de la distribución de probabilidad triangular, de la variable continua aleatoria x. (Véase la gráfica marginal).
- **a.** Calcule el valor de la constante c.
- **a.** Calcule la probabilidad de que x esté entre 0 y 3.
- **a.** Calcule la probabilidad de que x esté entre 2 y 9.

5-3 Aplicaciones de las distribuciones normales

Todos los ejemplos y ejercicios de la sección 5-2 son poco realistas, ya que incluyeron la distribución normal *estándar* (con una media de 0 y una desviación estándar de 1). En esta sección incluimos distribuciones normales no estándar, de modo que podamos trabajar con aplicaciones reales y prácticas. Sin embargo, es posible transformar valores de una distribución normal no estándar a una distribución normal estándar, para así continuar utilizando los mismos procedimientos de la sección 5-2.

> **Si convertimos valores en puntuaciones estándar, empleando la fórmula 5-2, entonces los procedimientos para trabajar con todas las distribuciones normales son los mismos que los de la distribución normal estándar.**

Fórmula 5-2 $z = \dfrac{x - \mu}{\sigma}$ (redondear puntuaciones z hasta dos decimales)

El uso continuo de la tabla A-2 requiere la comprensión y la aplicación del principio anterior. (Si utiliza ciertas calculadoras o programas de cómputo, no será necesaria la transformación a puntuaciones z, ya que las probabilidades se calculan de manera directa). Sin importar el método que utilice, debe comprender con claridad el principio básico anterior, puesto que constituye un fundamento importante de los conceptos que se introducen en los siguientes capítulos.

La figura 5-2 ilustra la transformación de una distribución no estándar a una estándar. El área de cualquier distribución normal que se limita por alguna puntuación x (como en la figura 5-12a), es igual que el área que se limita por la puntuación z equivalente en la distribución normal estándar (como en la figura 5-12b). Lo anterior significa que cuando se trabaja con una distribución normal no estándar, a veces se utiliza la tabla A-2 de la misma forma que se empleó en la sección 5-2, siempre y cuando los valores se conviertan primero a puntuaciones z. Cuando calcule áreas en una distribución normal no estándar, utilice este procedimiento:

1. Dibuje una curva normal, etiquete la media y los valores específicos de x; después, *sombree* la región que representa la probabilidad que se desea.
2. Para cada valor relevante de x que sea un límite de la región que se sombreó, utilice la fórmula 5-2 para transformar el valor a la puntuación z equivalente.

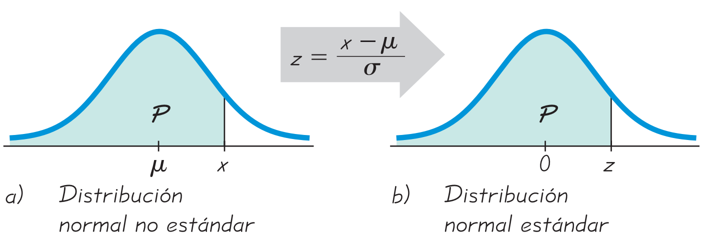


FIGURA 5-12 Transformación de una distribución normal no estándar a una distribución normal estándar

3. Remítase a la tabla A-2 para encontrar el área de la región que se sombreó, la cual constituye la probabilidad que se desea.

El siguiente ejemplo aplica estos tres pasos e ilustra la relación entre una distribución no normal típica y la distribución normal estándar.

EJEMPLO Diseño de automóviles La altura, en posición de sentado (del asiento a la cima de la cabeza), de los conductores debe tomarse en cuenta en el diseño de un nuevo modelo de automóvil. Los hombres tienen alturas que se distribuyen normalmente, con una media de 36.0 pulgadas y una desviación estándar de 1.4 pulgadas (según datos de una encuesta antropométrica de Gordon, Clauser *et al.*). Los ingenieros elaboran planes que pueden acomodar a hombres con alturas, estando sentados, de hasta 38.8 pulgadas, pero aquellos con mayor altura no se ajustan. Si se selecciona un hombre aleatoriamente, calcule la probabilidad de que su altura, estando sentado, sea menor que 38.8 pulgadas. Con base en ese resultado, ¿es factible el actual diseño de ingeniería?

SOLUCIÓN

Paso 1: Observe la figura 5-13, donde está marcada la media de 36.0 y la altura máxima de un hombre sentado de 38.8 pulgadas, en la cual el área que representa la probabilidad que se busca se sombreó. (Continuamos utilizando la misma correspondencia entre *probabilidad* y *área*, tal como se introdujo la sección 5-2).

Paso 2: Para usar la tabla A-2, primero hay que aplicar la fórmula 5-2, para convertir la distribución de alturas a una distribución normal estándar. La altura de 38.8 pulgadas se convierte en una puntuación z, de la siguiente manera:

$$z = \frac{x - \mu}{\sigma} = \frac{38.8 - 36.0}{1.4} = 2.00$$

Tal resultado demuestra que la altura de un hombre sentado, de 38.8 pulgadas, se encuentra por arriba de la media de 36.0 pulgadas por 2.00 desviaciones estándar.

Paso 3: Remitiéndonos a la tabla A-2, encontramos que $z = 2.00$ corresponde a un área de 0.9772.

INTERPRETACIÓN Hay una probabilidad de 0.9772 de seleccionar aleatoriamente a un hombre que, sentado, tenga una altura menor que 38.8 pulgadas. En símbolos, esto se expresa como

$$P(x < 38.8 \text{ pulg.}) = P(z < 2.00) = 0.9772$$

continúa

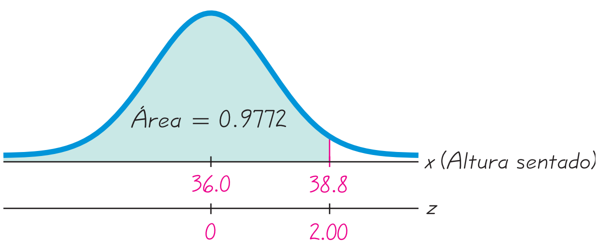


FIGURA 5-13 Distribución normal de la altura de hombres sentados

Atajo para ensayo clínico

¿Qué haría si estuviera probando un tratamiento y, antes de que su estudio termine, se da cuenta de que es claramente efectivo? Debería acortar el estudio e informar a todos los participantes acerca de la efectividad del tratamiento. Lo anterior fue lo que sucedió cuando se probó la hidroxiurea como tratamiento para la anemia falciforme. El estudio se programó para durar cerca de 40 meses, pero la efectividad del tratamiento se hizo evidente y el estudio se detuvo después de 36 meses. (Véase "Trial Halted as Sickle Cell Treatment Proves Itself" de Charles Marwick, *Journal of the American Medical Association,* vol. 273, núm. 8).

Otra forma de interpretar este resultado es concluir que el 97.72% de los hombres tienen alturas menores que 38.8 pulgadas, cuando están sentados en un automóvil. Una consecuencia importante de tal resultado es que el 2.28% de los hombres no se ajustan al automóvil. El fabricante debe decidir ahora si puede costear la pérdida del 2.28% de los conductores de automóviles varones.

PROBLEMA DEL CAPÍTULO 5

EJEMPLO Asientos de expulsión de aviones de propulsión a chorro En el problema del capítulo se señaló que la Fuerza Aérea de Estados Unidos estuvo usando los asientos expulsión ACES-II, que se diseñaron para hombres con un peso entre 140 y 211 libras. Siendo que los pesos de mujeres se distribuyen normalmente, con una media de 143 libras y una desviación estándar de 29 libras (según datos del National Health Survey), ¿qué porcentaje de las mujeres tiene pesos que se encuentran dentro de dichos límites?

SOLUCIÓN Observe la figura 5-14, que muestra la región que se sombreó de las mujeres que pesan entre 140 y 211 libras. Es posible encontrar esa área que se sombreó directamente de la tabla A-2, y obtenerla de manera indirecta utilizando los procedimientos básicos presentados en la sección 5-2. Para esto, primero debemos encontrar el área acumulativa de la izquierda, hasta 140 libras, y el área acumulativa de la izquierda, hasta 211 libras; después, se obtiene la diferencia entre ambas áreas.

Obtención del área acumulativa hasta 140 libras:

$$z = \frac{x - \mu}{\sigma} = \frac{140 - 143}{29} = -0.10$$

Si usamos la tabla A-2, encontraremos que $z = -0.10$ corresponde a un área de 0.4602, como se aprecia en la figura 5-14.

Obtención del área acumulativa hasta 211 libras:

$$z = \frac{x - \mu}{\sigma} = \frac{211 - 143}{29} = 2.34$$

Si usamos la tabla A-2, encontraremos que $z = 2.34$ corresponde a un área de 0.9904, como se observa en la figura 5-14.

Obtención del área entre 140 libras y 211 libras:

$$\text{Área sombreada} = 0.9904 - 0.4602 = 0.5302$$

INTERPRETACIÓN Encontramos que el 53.02% de las mujeres tienen pesos que se encuentran entre los límites del asiento de expulsión de 140 y 211 libras. Lo anterior significa que el 46.98% de las mujeres no tienen pesos que estén dentro de los límites actuales; muchas mujeres pilotos correrían el riesgo de dañarse gravemente si tuvieran que utilizar el asiento de expulsión.

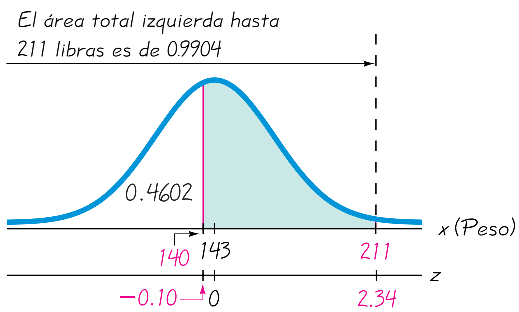


FIGURA 5-14 Pesos de mujeres y límites de los asientos de expulsión

Filas

La teoría de las filas es una rama de las matemáticas que se apoya en la probabilidad y la estadística. El estudio de las filas o líneas de espera es importante para negocios como supermercados, bancos, restaurantes de comida rápida, líneas aéreas y parques de diversiones. Los supermercados Grand Union tratan de mantener filas en las cajas de no más de tres compradores. Wendy's introdujo el sistema "Express Pak" para agilizar el servicio a los numerosos clientes que atienden en sus automóviles. Disney realiza extensos estudios de filas en sus parques de diversiones, para mantener contentos a sus visitantes y planear sus expansiones. Los laboratorios Bell aplican la teoría de las filas para optimizar el uso de las redes telefónicas, en tanto que las fábricas la emplean para diseñar líneas de producción eficientes.

Cálculo de valores de áreas conocidas

Los ejemplos anteriores en esta sección son del mismo tipo: se nos dan valores de límites específicos y debemos encontrar un área (o probabilidad o porcentaje). En muchos casos prácticos y reales, ya se conoce el área (o probabilidad o porcentaje), pero habrá que encontrar el (los) valor(es) relevante(s). Cuando busque valores de áreas conocidas, asegúrese de considerar lo siguiente:

1. *No confunda las puntuaciones z y las áreas.* Recuerde que las puntuaciones z son *distancias* a lo largo de la escala horizontal, en tanto que las áreas son *regiones* bajo la curva normal. La tabla A-2 lista puntuaciones z en las columnas de la izquierda y a lo largo del renglón superior, pero las áreas se localizan en el cuerpo de la tabla.
2. *Elija el lado correcto de la gráfica (derecho/izquierdo).* Un valor que separa el 10% superior del resto se localizará en el lado derecho de la gráfica, pero un valor que separa el 10% inferior se ubicará en el lado izquierdo de la gráfica.
3. Una puntuación z debe ser *negativa* siempre que esté localizada en la mitad *izquierda* de la distribución normal.
4. Las áreas (o probabilidades) son positivas o tienen valores de 0, pero nunca son negativas.

Las gráficas son sumamente útiles para visualizar, comprender y trabajar con éxito con las distribuciones de probabilidad normal; por lo tanto, deben emplearse siempre que sea posible.

Procedimiento del cálculo de valores con el uso de la tabla A-2 y la fórmula 5-2

1. Dibuje una curva de distribución normal, anote la probabilidad o porcentaje dados en la región apropiada de la gráfica e identifique el (los) valor(es) que se busca(n).
2. Utilice la tabla A-2 para encontrar la puntuación z correspondiente al área izquierda acumulativa, limitada por x. Remítase al *cuerpo* de la tabla A-2 para localizar el área más cercana; después, identifique la puntuación z correspondiente.
3. Para emplear la fórmula 5-2, sustituya los valores de μ, σ y la puntuación z que se obtuvo en el paso 2; después, calcule x. Con base en el formato de la fórmula 5-2, calculamos x de la siguiente manera:

$$x = \mu + (\underset{\uparrow}{z} \cdot \sigma) \qquad \text{(otra forma de la fórmula 5-2)}$$

(Si z se localiza a la izquierda de la media, asegúrese de que sea un número negativo).

4. Remítase al dibujo de la curva para verificar que la solución es lógica en el contexto de la gráfica y en el del problema.

El siguiente ejemplo utiliza el procedimiento que se acaba de describir.

PROBLEMA DEL CAPÍTULO 5

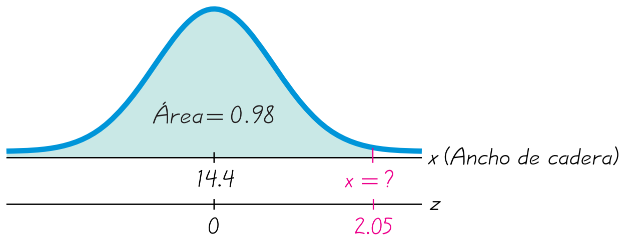
EJEMPLO Ancho de caderas y asientos de aviones Al diseñar los asientos que se habrán de instalar en aviones comerciales, los ingenieros buscan hacerlos con una anchura suficiente para que quepa el 98% de los hombres. (Acomodar al 100% de los hombres requeriría asientos muy anchos, que serían muy caros). El ancho de las caderas de los hombres se distribuye de manera normal, con una media de 14.4 pulgadas y

continúa

El medio de una encuesta puede afectar los resultados

En una encuesta que se realizó a individuos católicos de Boston, se les preguntó si pensaban que debían proporcionarse anticonceptivos a las mujeres solteras. En entrevistas personales, el 44% de los sujetos respondieron afirmativamente. Pero en un grupo similar, que se contactó por correo o por teléfono, el 75% de las personas respondió que sí a la misma pregunta.

FIGURA 5-15 Distribución del ancho de cadera de hombres



una desviación estándar de 1.0 pulgadas (de acuerdo con datos de una encuesta antropométrica de Gordon, Clauser *et al*.). Calcule P_{98}, es decir, obtenga el ancho de cadera de hombres que separa al 98% inferior del 2% superior.

SOLUCIÓN

Paso 1: Iniciamos con la gráfica de la figura 5-15. Ya anotamos la media de 14.4 pulgadas, sombreamos el área que representa al 98% inferior e identificamos el valor que se desea como x.

Paso 2: En el *cuerpo* de la tabla A-2 buscamos un área de 0.9800. (El área de 0.98 que se muestra en la figura 5-15 es un área acumulativa de la izquierda y exactamente el tipo de área que se lista en la tabla A-2). El área más cercana a 0.98, que es 0.9798, corresponde a la puntuación z de 2.05.

Paso 3: Con $z = 2.05$, $\mu = 14.4$ y $\sigma = 1.0$, calculamos x empleando la fórmula 5-2 de manera directa o utilizando la siguiente versión de la fórmula 5-2:

$$x = \mu + (z \cdot \sigma) = 14.4 + (2.05 \cdot 1.0) = 16.45$$

Paso 4: Si permitimos que $x = 16.45$ en la figura 5-15, veremos que esta solución es razonable, ya que el percentil 98 debe ser mayor que la media de 14.4.

INTERPRETACIÓN El ancho de cadera de 16.5 pulgadas (que se redondeó a un decimal, como en μ y σ) separa al 98% inferior del 2% superior. Es decir, los asientos que se diseñan para un ancho de cadera de hasta 16.5 pulgadas se ajustarán al 98% de los hombres. Este tipo de análisis se utiliza para diseñar los asientos que se emplean actualmente en los aviones comerciales.

EJEMPLO Diseño de tableros de automóviles Al diseñar la ubicación de un reproductor de CD en un nuevo modelo de automóviles, los ingenieros deben considerar el alcance frontal del conductor. Si el reproductor de CD se coloca más allá del alcance, el conductor tiene que mover su cuerpo de manera que podría distraerse, lo cual sería peligroso. (No desearíamos que alguien se lastimara tratando de escuchar lo mejor de Barry Manilow). Los diseñadores deciden que el reproductor debe ubicarse de manera que esté dentro del alcance del 95% de las mujeres. Las mujeres tienen alcances frontales distribuidos normalmente, con una media de 27.0 pulgadas y una desviación estándar de 1.3 pulgadas (según datos de una encuesta antropométrica de Gordon, Churchill *et al*.). Calcule el alcance frontal de las mujeres que separa al 95% superior del resto.

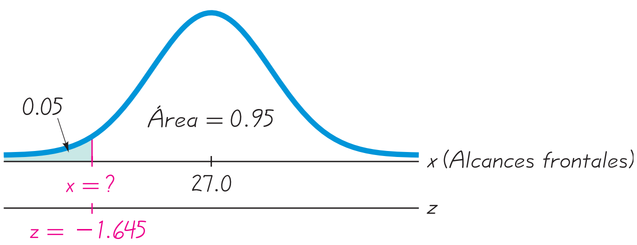


FIGURA 5-16 Cálculo del valor que separa al 95% superior

SOLUCIÓN

Paso 1: Iniciamos con la gráfica de la figura 5-16. Ya incluimos la media de 27.0 pulgadas e identificamos el área que representa el 95% superior de los alcances frontales. Aun cuando el problema se refiere al 95% superior, la tabla A-2 requiere que trabajemos con un área *izquierda* acumulativa, por lo que restamos 0.95 de 1 para obtener 0.05, que aparece como la región que se sombreó.

Paso 2: En el *cuerpo* de la tabla A-2 buscamos un área de 0.05. Las áreas más cercanas a 0.05 son 0.0505 y 0.0495, pero hay un asterisco que indica que un área de 0.05 corresponde a una puntuación z de -1.645.

Paso 3: Con $z = -1.645$, $\mu = 27.0$ y $\sigma = 1.3$, calculamos x empleando la fórmula 5-2 de manera directa o utilizando la siguiente versión de la fórmula 5-2:

$$x = \mu + (z \cdot \sigma) = 27.0 + (-1.645 \cdot 1.3) = 24.8615$$

Paso 4: Si permitimos que $x = 24.8615$ en la figura 5-16, veremos que tal solución es razonable, ya que el alcance frontal que separa al 95% superior del 5% inferior debe ser menor que la media de 27.0 pulgadas.

INTERPRETACIÓN El alcance frontal de 24.9 pulgadas (redondeado) separa al 95% superior del resto, ya que el 95% de las mujeres tienen alcances frontales mayores que 24.9 pulgadas y el 5% tienen alcances frontales menores que 24.9 pulgadas.

Utilizando la tecnología

STATDISK Seleccione **Analysis, Probability Distributions, Normal Distribution**; introduzca los valores de la media y de la desviación estándar; después, deslice el mouse a la derecha o a la izquierda hasta obtener el valor que se desea. Puede lograr mayor precisión si emplea el mouse para arrastrar parte de la curva, de manera que pueda amplificarla.

Minitab

- Para encontrar el área acumulativa que está a la izquierda de una puntuación z (como en la tabla A-2), seleccione **Calc, Probability Distributions, Normal, Cumulative probabilities;** introduzca la media y la desviación estándar; después, haga clic en el botón de **Input Constant,** e introduzca el valor.
- Para encontrar un valor correspondiente a un área que se conoce, seleccione **Calc, Probability Distributions, Normal;** después, **Inverse cumulative probabilities,** e introduzca la media y la desviación estándar. Elija la opción **Input constant** e introduzca el área total que se encuentra a la izquierda del valor dado.

Excel

- Para encontrar el área acumulativa a la izquierda de un valor (como en la tabla A-2), haga clic en ***fx***; después, seleccione **Statistical, NORMDIST.** En el cuadro de diálogo, introduzca el valor de x, la media y la desviación estándar; finalmente, 1 en el espacio "cumulative".

continúa

- Para encontrar el valor correspondiente a un área que se conoce, seleccione ***fx*, Statistical, NORMINV;** ahora proceda a introducir la información en el cuadro de diálogo. Cuando anote el valor de probabilidad, introduzca el área total a la izquierda del valor dado.

TI-83 Plus

- Para calcular el área entre dos valores, presione **2nd VARS, 2**(para normalcdf); después, proceda a introducir los dos valores, la media y la desviación estándar, todos separados por comas (como en valor izquierdo, valor derecho, media, desviación estándar).
- Para encontrar un valor correspondiente a un área que se conoce, presione **2nd VARS, 3** (para invNorm); ahora proceda a introducir el área total a la izquierda del valor, la media y la desviación estándar con el formato área total a la izquierda, media, desviación estándar, incluyendo las comas.

5-3 Destrezas y conceptos básicos

***Puntuaciones de CI**. En los ejercicios 1 a 8, suponga que sujetos adultos tienen puntuaciones de CI que se distribuyen normalmente, con una media de 100 y una desviación estándar de 15 (como en la prueba Weschler). (Sugerencia: Dibuje una gráfica en cada caso).*

1. Calcule la probabilidad de que un adulto que se seleccione al azar tenga un CI menor de 115.

2. Calcule la probabilidad de que un adulto que se seleccione al azar tenga un CI mayor de 131.5 (requisito para ser miembro de la organización Mensa).

3. Calcule la probabilidad de que un adulto que se seleccione aleatoriamente tenga un CI entre 90 y 110 (denominado rango *normal*).

4. Calcule la probabilidad de que un adulto seleccionado aleatoriamente tenga un CI entre 110 y 120 (denominado *normal brillante*).

5. Calcule P_{20}, que es la puntuación de CI que separa al 20% inferior del 80% superior.

6. Calcule P_{80}, que es la puntuación de CI que separa al 80% inferior del 20% superior.

7. Calcule la puntuación de CI que separa al 15% superior del resto.

8. Calcule la puntuación de CI que separa al 55% superior del resto.

9. Temperaturas corporales Con base en los resultados muestrales del conjunto de datos 4 del Apéndice B, suponga que las temperaturas corporales humanas se distribuyen normalmente, con una media de 98.20°F y una desviación estándar de 0.62°F.

a. El hospital Bellevue, en la ciudad de Nueva York, establece que la temperatura más baja que se considera como fiebre es de 100.6°F. ¿Qué porcentaje de personas normales y saludables se consideraría que tiene fiebre? ¿Sugiere este porcentaje que un punto de corte de 100.6°F es apropiado?

b. Los médicos desean seleccionar una temperatura mínima como requisito para aplicar mayor cantidad de exámenes médicos. ¿Qué temperatura debe ser si deseamos que sólo el 5.0% de las personas saludables la excedan? (Un resultado como éste es un *falso positivo*, lo que significa que el resultado de la prueba es positivo, pero el sujeto no está realmente enfermo).

10. Duración de embarazos La duración de los embarazos se distribuye normalmente, con una media de 268 días y una desviación estándar de 15 días.

a. Un uso clásico de la distribución normal se inspiró en una carta dirigida a “Dear Abby”, en la que una mujer afirmaba haber dado a luz 308 días después de una breve visita de su esposo, que trabajaba en la Marina. Con esta información, calcule la probabilidad de que un embarazo dure 308 días o más. ¿Qué sugiere el resultado?

b. Si estipulamos que un bebé es *prematuro* cuando la duración del embarazo se encuentra en el 4% inferior, calcule la duración que separa a los bebés prematuros de aquellos que no lo son. Los bebés prematuros suelen requerir cuidados especiales y este resultado sería útil para que los administradores de hospitales planeen dichos cuidados.

11. Requisitos de la prueba SAT La combinación de las calificaciones verbales y de matemáticas de mujeres que toman la prueba SAT-I se distribuye de manera normal, con una media de 998 y una desviación estándar de 202 (de acuerdo con datos del College Board). El College of Westport incluye una calificación mínima de 1100 entre sus requisitos.

a. ¿Qué porcentaje de mujeres *no* satisfacen este requisito?

b. Si se cambia el requisito a "una calificación que esté dentro del 40% superior", ¿cuál es la calificación mínima que se requiere? ¿Qué dificultad práctica se crearía si se anunciara que el nuevo requisito es ubicarse en "el 40% superior"?

12. Diseño de cascos Los ingenieros deben tomar en cuenta la anchura de las cabezas de los hombres cuando diseñan cascos para motociclistas. La anchura de las cabezas de los hombres se distribuye normalmente, con una media de 6.0 pulgadas y una desviación estándar de 1.0 pulgadas (según datos de una encuesta antropométrica de Gordon, Churchill *et al*.). Por limitaciones económicas, los cascos se diseñarán para que se ajusten a todos los hombres, excepto al 2.5% con anchuras menores y al 2.5% con anchuras más grandes. Calcule las anchuras de cabeza mínima y máxima que se ajustarán a los cascos.

13. Garantía de televisores El tiempo de reemplazo de los televisores se distribuye normalmente, con una media de 8.2 años y una desviación estándar de 1.1 años (de acuerdo con datos de "Getting Things Fixed", *Consumer Reports*).

a. Calcule la probabilidad de que un televisor que se selecciona aleatoriamente tenga un tiempo de reemplazo menor de 5.0 años.

b. Si usted desea ofrecer una garantía tal que sólo el 1% de los televisores se reemplacen antes de que expire la garantía, ¿cuál debe ser la duración de la garantía?

14. Garantía de reproductores de CD El tiempo de reemplazo de los reproductores de CD se distribuye normalmente, con una media de 7.1 años y una desviación estándar de 1.4 años (con base en datos de "Getting Things Fixed", *Consumer Reports*).

a. Calcule la probabilidad de que un reproductor de CD, que se seleccionó aleatoriamente, tenga un tiempo de reemplazo menor de 8.0 años.

b. Si usted desea ofrecer una garantía tal que sólo el 2% de los reproductores se reemplace antes de que expire la garantía, ¿cuál debe ser la duración de la garantía?

15. M&M A continuación se presentan los resultados de Minitab de los pesos (en gramos) de los 100 dulces M&M listados en el conjunto de datos 19 del Apéndice B. Aun cuando la media y la desviación estándar son estadísticos muestrales, suponga que son los parámetros poblacionales de todos los M&M.

a. Calcule el porcentaje de pesos menores que 8.88925 g. ¿De qué forma coincide el resultado con el valor de 0.88925 presentado como Q_1, el primer cuartil?

b. Calcule el valor de Q_1. ¿De qué forma coincide el resultado con el valor de 0.88925 que se muestra en la representación visual?

```
Variable       N     Media   Mediana Desv. Est.  Desv. Est.  SE Mean
M&M          100   0.91470   0.91050    0.91307     0.03691  0.00369

Variable  Mínimo    Máximo        Q1          Q3
M&M      0.83800   1.03300   0.88925     0.93375
```

TI-83 Plus

```
1-Var Stats
 x̄=.8168222222
 Σx=29.4056
 Σx²=24.0211202
 Sx=.007507372
 σx=.0074023687
↓n=36
```

16. Pesos de Coca Cola regular Al margen se presenta la representación visual de la calculadora TI-83 Plus con los pesos (en libras) de la Coca Cola regular, tal como se listan en el conjunto de datos 17 del Apéndice B. Aun cuando la media y la desviación estándar son estadísticos muestrales, suponga que son parámetros poblacionales de todas las latas de Coca Cola regular. (Utilice el valor de Sx para la desviación estándar).

a. Si se selecciona aleatoriamente una lata de Coca Cola regular, calcule la probabilidad de que su contenido pese más de 0.8250 libras.

b. Con el propósito de verificar la producción futura de Coca Cola, calcule los pesos que separan al 2.5% inferior y al 2.5% superior.

Estaturas de mujeres. *En los ejercicios 17 a 20, suponga que la estatura de las mujeres se distribuye de forma normal, con una media dada $\mu = 63.6$ pulgadas y una desviación estándar dada $\sigma = 2.5$ pulgadas (de acuerdo con datos del National Health Survey). En cada caso, dibuje una gráfica.*

17. Estatura requerida por el Club Beanstalk El Club Beanstalk, una organización social para personas altas, requiere que las mujeres midan al menos 70 pulgadas (o 5 pies 10 pulgadas). ¿Qué porcentaje de las mujeres cumple este requisito?

18. Estaturas requeridas para mujeres soldados El ejército de Estados Unidos requiere que las mujeres midan entre 58 y 80 pulgadas. Calcule el porcentaje de mujeres que cumplen este requisito. ¿Se les negará a muchas mujeres la oportunidad de unirse al ejército porque son muy altas o muy bajas?

19. Estaturas requeridas para las bailarinas *Rockettes* Para estar en una compañía de baile con una apariencia uniforme, las famosas bailarinas *Rockette*, del Radio City Music Hall de Nueva York, deben sujetarse a ciertas restricciones de estaturas. Como las mujeres ahora son más altas, un cambio reciente requiere que una bailarina *Rockette* tenga una estatura entre 66.5 y 71.5 pulgadas. Si se selecciona una mujer al azar, ¿cuál es la probabilidad de que cumpla este nuevo requisito de estatura? ¿Qué porcentaje de mujeres cumplen este nuevo requisito de estatura? ¿Parecería que las *Rockettes* son generalmente más altas que la mujer común?

20. Estaturas requeridas para las bailarinas *Rockettes* El ejercicio 19 identificó requisitos de estatura específicos para las bailarinas *Rockettes*. Suponga que dichos requisitos deben cambiarse porque muy pocas mujeres los cumplen. ¿Cuáles serían las nuevas estaturas mínima y máxima que se permiten si se excluyera al 20% de mujeres más bajas y al 20% de mujeres más altas?

5-3 Más allá de lo básico

21. Unidades de medición Los pesos de las mujeres se distribuyen normalmente, con una media de 143 libras y una desviación estándar de 29 libras.

a. Si los pesos de mujeres individuales se expresan en libras, ¿cuáles serían las unidades que se utilizarían para las puntuaciones z correspondientes a los pesos individuales?

b. Si los pesos de todas las mujeres se convierten a puntuaciones z, ¿cuál es la media, la desviación estándar y la distribución de estas puntuaciones z?

c. ¿Cuál es la distribución, la media y la desviación estándar de los pesos de las mujeres después de convertirlos a kilogramos (1 libra = 0.4536 kg)?

22. Uso de la corrección por continuidad Hay muchas situaciones en las que una distribución normal se utiliza como una buena aproximación de una variable aleatoria con sólo valores *discretos*. En tales casos, podemos emplear esta *corrección por continuidad*: represente cada número entero con el intervalo que va desde 0.5 por debajo del número hasta 0.5 por arriba de él. Suponga que las puntuaciones de CI son todas números enteros con una distribución aproximadamente normal, con una media de 100 y una desviación estándar de 15.

a. Sin utilizar la corrección por continuidad, calcule la probabilidad de seleccionar aleatoriamente a alguien con una puntuación de CI mayor que 105.

b. Utilice la corrección por continuidad y calcule la probabilidad de seleccionar aleatoriamente a alguien con una puntuación de CI mayor que 105.

c. Compare los resultados de los incisos *a* y *b*.

23. Normalización de calificaciones de un examen Una profesora informa a sus alumnos de la clase de psicología que un examen es muy difícil, pero que las calificaciones se normalizarán. Las calificaciones de este examen se distribuyen normalmente, con una media de 25 y una desviación estándar de 5.

a. Si las normaliza sumando 50 a cada calificación, ¿cuál es la nueva media? ¿Cuál es la nueva desviación estándar?

b. ¿Será justo normalizarlas sumando 50 a cada calificación? ¿Por qué?

c. Si las calificaciones se normalizan según el siguiente esquema (en lugar de sumar 50), calcule los límites numéricos de cada calificación.

A: 10% superior

B: Calificaciones por arriba del 70% inferior y por debajo del 10% superior

C: Calificaciones por arriba del 30% inferior y por debajo del 30% superior

D: Calificaciones por arriba del 10% inferior y por debajo del 70% superior

F: 10% inferior

d. ¿Cuál método de normalización de las calificaciones es más justo: sumar 50 a cada calificación o emplear el esquema del inciso *c*? Explique.

24. Calificaciones del SAT Según datos del College Entrance Examination Board, las calificaciones de la prueba SAT-I tienen una media de 1017 y Q_1 es 880, con una distribución aproximadamente normal. Calcule la desviación estándar y después utilice este resultado para obtener P_{99}.

25. Pruebas SAT y ACT Las calificaciones de mujeres en la prueba SAT-I se distribuyen de manera normal, con una media de 998 y una desviación estándar de 202. Las calificaciones de mujeres en la prueba ACT se distribuyen de manera normal, con una media de 20.9 y una desviación estándar de 4.6. Suponga que las dos pruebas emplean escalas distintas para medir la misma habilidad.

a. Si una mujer obtiene una calificación en el SAT que corresponde al percentil 67, calcule su calificación real en el SAT y su calificación equivalente en el ACT.

b. Si una mujer obtiene una calificación de 1220 en el SAT, calcule su calificación equivalente en el ACT.

5-4 Distribuciones muestrales y estimadores

Comenzamos a embarcarnos en un viaje que nos permitirá conocer las poblaciones al obtener datos de muestras. Las secciones 5-5 y 5-6 proporcionan conceptos importantes que revelan el comportamiento de medias de muestra y proporciones de muestra. Antes de considerar dichos conceptos, observemos el comportamiento de los estadísticos muestrales en general. El principal objetivo de esta sección es aprender lo que conocemos como la *distribución muestral de un estadístico*; otro objetivo importante es aprender un principio básico acerca de la distribución muestral de medias de muestra y la distribución muestral de proporciones de muestra.

Iniciemos con las medias de muestra. En lugar de ser muy abstractos, consideremos la *población* que consiste en los valores 1, 2, 5. El McGwire Electronics Center estuvo abierto sólo durante tres días, por un personal de ventas descortés, a un pobre plan de negocios, publicidad ineficaz y a una ubicación inadecuada. Durante el primer día se vendió un teléfono celular, durante el segundo día dos teléfonos celulares y sólo cinco durante el tercer día. Puesto que 1, 2, 5 constituyen la población completa, es fácil calcular los valores de los parámetros poblacionales: $\mu = 2.7$ y $\sigma = 1.7$. Para calcular el valor de la desviación estándar poblacional σ, empleamos

$$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}} \text{ en lugar de } s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$

Es raro que conozcamos todos los valores de una población completa. Lo más común es que exista una población grande desconocida que queremos investigar. Como no es práctico encuestar a cada miembro de la población, obtenemos una muestra; luego, con base en las características de ésta, hacemos estimaciones acerca de las características de la población. Por ejemplo, la Hartford Insurance Company querría conocer la población de las edades de todos los conductores, por medio de la obtención de una muestra de dichas edades.

Ya que los valores 1, 2, 5 constituyen una población completa, consideremos muestras de tamaño 2. Con sólo tres valores poblacionales, hay únicamente nueve posibles muestras diferentes de tamaño 2, suponiendo que el muestreo se realiza con reemplazo. Es decir, se reemplaza cada valor seleccionado antes de realizar una nueva selección.

¿Por qué se hace el muestreo *con* reemplazo? Para muestras pequeñas como la que estudiamos, el muestreo *sin reemplazo* tiene la ventaja práctica de evitar una duplicación inútil, siempre que se selecciona el mismo elemento más de una vez. Sin embargo, a lo largo de la presente sección nos interesamos particularmente en el muestreo *con reemplazo,* por las siguientes razones: **1.** Cuando se selecciona una muestra relativamente pequeña, de una población grande, no hay mucha diferencia si realizamos la muestra con reemplazo o sin él. **2.** El muestreo con reemplazo da como resultado sucesos independientes que no se afectan por resultados previos; asimismo, los sucesos independientes son más fáciles de analizar y derivan en fórmulas más simples. Por eso nos enfocamos en el comportamiento de muestras que se seleccionan aleatoriamente *con reemplazo*.

Cuando tomamos una muestra de dos valores con reemplazo, de la población de 1, 2, 5, cada una de las nueve muestras es igualmente posible, con una probabilidad de $1/9$. La tabla 5-2 lista las nueve muestras posibles de tamaño 2, junto con los estadísticos para cada muestra. Esta tabla contiene mucha información, pero consideremos primero la columna de medias de muestra. Puesto que se listan todos los posibles valores de $\bar{x}$,y puesto que se sabe que la probabilidad de cada uno es de $1/9$, tenemos una distribución de probabilidad. (Recuerde, una distribución de probabilidad describe la probabilidad de cada valor de una variable aleatoria, en tanto que la variable aleatoria, en este caso, es el valor de la media muestral). Como varios métodos importantes de estadística inician con una media muestral, que se utiliza subsecuentemente para hacer inferencias acerca de la media poblacional, es importante comprender el comportamiento de tales medias de muestra. Otros métodos de estadística importantes comienzan con una proporción de muestra que se emplea subsecuentemente para hacer inferencias acerca de la proporción

Tabla 5-2 Distribuciones muestrales de diferentes estadísticos (para muestras de tamaño 2, obtenidas con reemplazo de la población 1, 2, 5)

Muestra	Media $\bar{x}$	Mediana	Rango	Varianza s^2	Desviación estándar s	Proporción de números impares	Probabilidad
1, 1	1.0	1.0	0	0.0	0.000	1	1/9
1, 2	1.5	1.5	1	0.5	0.707	0.5	1/9
1, 5	3.0	3.0	4	8.0	2.828	1	1/9
2, 1	1.5	1.5	1	0.5	0.707	0.5	1/9
2, 2	2.0	2.0	0	0.0	0.000	0	1/9
2, 5	3.5	3.5	3	4.5	2.121	0.5	1/9
5, 1	3.0	3.0	4	8.0	2.828	1	1/9
5, 2	3.5	3.5	3	4.5	2.121	0.5	1/9
5, 5	5.0	5.0	0	0.0	0.000	1	1/9
Media de valores de los estadísticos	2.7	2.7	1.8	2.9	1.3	0.667	
Parámetro poblacional	2.7	2	4	2.9	1.7	0.667	
¿Coincide el estadístico muestral con el parámetro poblacional?	Sí	No	No	Sí	No	Sí	

poblacional, así que es importante comprender el comportamiento de estas proporciones de muestra. En general, es importante entender el comportamiento de los estadísticos muestrales. El "comportamiento" de un estadístico se puede conocer al comprender su distribución.

Definición

La **distribución muestral de la media** es la distribución de probabilidad de medias muestrales, donde todas las muestras tienen el mismo tamaño *n*. (En general, la distribución de muestreo de cualquier estadístico es la distribución de probabilidad de dicho estadístico).

EJEMPLO Distribución muestral de la media Una *población* se compone de los valores 1, 2, 5, en tanto que la tabla 5-2 incluye todas las distintas muestras posibles del tamaño $n = 2$. La probabilidad de cada muestra se lista en la tabla 5-2 como $1/9$. Identifique la distribución muestral específica de la media de las muestras de tamaño $n = 2$, que se seleccionan aleatoriamente con reemplazo, de la población 1, 2, 5. También, calcule la media de esta distribución muestral. ¿Coinciden las medias de muestra con el valor de la media poblacional?

continúa

Tabla 5-3
Distribución muestral de la media

Media $\bar{x}$	Probabilidad
1.0	1/9
1.5	1/9
3.0	1/9
1.5	1/9
2.0	1/9
3.5	1/9
3.0	1/9
3.5	1/9
5.0	1/9

Esta tabla lista las medias de las muestras en la tabla 5-2, pero podría condensarse listando 1.0, 1.5, 2.0, 3.0, 3.5 y 5.0, junto con sus probabilidades correspondientes de 1/9, 2/9, 1/9, 2/9, 2/9 y 1/9.

SOLUCIÓN La distribución muestral de la media es la distribución de probabilidad que describe la probabilidad para cada valor de la media, y dichos valores se incluyen en la tabla 5-2. Así pues, la distribución muestral de la media se escribe usando la tabla 5-3. La media de la distribución muestral se calcula con dos métodos diferentes: **1.** utilice $\mu = \Sigma[x \cdot P(x)]$, que es la fórmula 4-2, o **2.** puesto que las nueve medias de muestra son igualmente posibles, podríamos simplemente calcular la media de esos nueve valores. Como la media poblacional también es 2.7, parece que las medias de muestra "coinciden" con el valor de la media poblacional, en lugar de subestimar o sobreestimar sistemáticamente la media poblacional.

En el ejemplo anterior, observamos que la media de todas las medias muestrales posibles es igual a la media de la población original, que es $\mu = 2.7$. Podemos generalizar esto como una propiedad de las medias de muestra: para un tamaño de muestras fijo, la media de todas las medias muestrales posibles es igual a la media de la población. Revisaremos esta importante propiedad en la siguiente sección, pero antes hagamos una observación evidente, aunque muy importante: *las medias de muestra varían*. Véase la tabla 5-3 y observe que las medias de muestra son diferentes. La primera media de muestra es 1.0, la segunda media de muestra es 1.5, etcétera. Esto nos conduce a la siguiente definición.

Definición

El valor de un estadístico, como la media muestral $\bar{x}$, depende de los valores particulares incluidos en la muestra, y generalmente varía de una muestra a otra. Tal variabilidad de un estadístico se denomina **variabilidad de muestreo**.

En el capítulo 2 estudiamos las características importantes de un conjunto de datos: centro, variación, distribución, datos distantes y patrón temporal (resumido con las siglas "CVDDT"). Al examinar las muestras en la tabla 5-2, ya identificamos una propiedad que describe el comportamiento de las medias de muestra: la media de medias de muestra es igual a la media de la población. Esta propiedad pone énfasis en la característica central; investigaremos otras características en la siguiente sección. Veremos que al incrementarse el tamaño de la muestra, la distribución muestral de medias de muestra tiende a convertirse en una *distribución normal*. (Esto no nos sorprende, ya que el título de este capítulo es "Distribuciones de probabilidad normal"). En consecuencia, la distribución normal tiene una importancia que va más allá de las aplicaciones que se ilustran en la sección 5-3. La distribución normal se utilizará en casos en los cuales deseamos emplear una media de muestra con el propósito de hacer alguna inferencia acerca de una media poblacional μ.

Distribución muestral de proporciones

Cuando hacemos inferencias acerca de una proporción poblacional, también es importante comprender el comportamiento de las proporciones muestrales. Definimos la distribución de proporciones muestrales de la siguiente manera.

Definición

Distribución muestral de la proporción: distribución de probabilidad de proporciones muestrales, donde todas las muestras tienen el mismo tamaño muestral *n*.

Uno de los usos clásicos de la estadística inferencial es el cálculo de alguna proporción muestral y su aplicación para hacer una inferencia acerca de la proporción poblacional. Encuestadores de la organización Gallup le preguntaron a 491 adultos que se seleccionaron al azar si estaban a favor de la pena de muerte para una persona que sentenciaron por homicidio. Los resultados mostraron que 319 individuos (o el 65% de ellos) se manifestaron a favor. El resultado muestral conduce a la inferencia de que "el 65% de todos los adultos están a favor de la pena de muerte para una persona sentenciada por homicidio". La proporción muestral de 319/491 se utilizó para estimar una proporción poblacional *p*, pero aprenderemos mucho más si comprendemos la distribución muestral de dichas proporciones muestrales.

EJEMPLO Distribución muestral de proporciones Una *población* se compone de los valores 1, 2, 5, en tanto que la tabla 5-2 incluye todas las distintas muestras posibles de tamaño $n = 2$, que se seleccionaron con reemplazo. Para cada muestra, considere la proporción de números *impares*. Identifique la distribución muestral para la proporción de números impares y después calcule su media. ¿Coinciden las proporciones muestrales con el valor de la proporción poblacional?

SOLUCIÓN Observe la tabla 5-2, donde las nueve proporciones muestrales que se indican son 1, 0.5, 1, 0.5, 0, 0.5, 1, 0.5, 1. Al combinar tales proporciones muestrales con sus probabilidades de 1/9 en cada caso, obtenemos la distribución muestral de proporciones que se resume en la tabla 5-4. La media de las proporciones *muestrales* es 0.667. Puesto que la población 1, 2, 5 contiene dos números impares, la proporción *poblacional* de números impares es también 2/3 o 0.667. En general, las proporciones muestrales tienden a coincidir con el valor de la proporción poblacional, y no a subestimar o sobreestimar sistemáticamente ese valor.

Tabla 5-4 Distribución muestral de proporciones

Proporción de números impares	Probabilidad
1	1/9
0.5	1/9
1	1/9
0.5	1/9
0	1/9
0.5	1/9
1	1/9
0.5	1/9
1	1/9

La tabla lista las proporciones de las muestras en la tabla 5-2, pero podría condensarse listando las proporciones de 0, 0.5 y 1, junto con sus probabilidades correspondientes de 1/9, 4/9 y 4/9.

El ejemplo anterior incluye una proporción bastante pequeña, de manera que ahora consideraremos los géneros de los senadores en el congreso 107o. Como sólo existen 100 miembros [13 mujeres (M) y 87 hombres (H)], listaremos la población completa:

H **M** H H **M** H H H H H H H **M** H H H H H H H
H H H H H H H H H H H H **M** **M** H H H H H H
H H H **M** H H H H H **M** H H H H H H H H H H
M H H H H H H H H H H H H H H H H **M** **M** **M**
H H H **M** H **M** H H H H H H H H H H H H H H

La proporción poblacional de senadoras es $p = 13/100 = 0.13$. Por lo general, no conocemos a todos los miembros de la población, por lo que debemos estimarla a

Tabla 5-5
Resultados de 100 muestras

Proporción de senadoras	Frecuencia
0.0	26
0.1	41
0.2	24
0.3	7
0.4	1
0.5	1
Media:	0.119
Desviación estándar:	0.100

partir de una muestra. Con el propósito de estudiar el comportamiento de las proporciones muestrales, listamos unas cuantas muestras de tamaño $n = 10$:

Muestra 1: H M H H M H H H H H $\rightarrow$ la proporción muestral es 0.2
Muestra 2: H M H H H H H H H H $\rightarrow$ la proporción muestral es 0.1
Muestra 3: H H H H H H M H H H $\rightarrow$ la proporción muestral es 0.1
Muestra 4: H H H H H H H H H H $\rightarrow$ la proporción muestral es 0
Muestra 5: H H H H H H H H M H $\rightarrow$ la proporción muestral es 0.1

Puesto que hay un número muy grande de muestras como éstas, no es posible listarlas todas. El autor seleccionó aleatoriamente 95 muestras adicionales, antes de detener las llantas de su automóvil. Si combinamos las 95 muestras adicionales con las cinco listadas antes, obtendremos las 100 muestras que se incluyen en la tabla 5-5.

Notamos, a partir de la tabla 5-5, que la media de las 100 proporciones muestrales es 0.119, pero si incluyéramos todas las otras posibles muestras de tamaño 10, la media de las proporciones muestrales sería igual a 0.13, que es el valor de la proporción poblacional. La forma de esa distribución se asemeja razonablemente a la que se obtendría con todas las muestras posibles de tamaño 10. Observamos que la distribución que se presenta en la figura 5-17 tiene cierto sesgo hacia la derecha, aunque con un poco de alargamiento se aproximaría a una distribución normal. En la figura 5-18 mostramos los resultados obtenidos a partir de 10,000 muestras, de tamaño 50, que se seleccionaron al azar y con reemplazo, de la lista anterior de 100 géneros. La figura 5-18 sugiere, con énfasis, que la distribución se aproxima a la forma de campana que caracteriza a una distribución normal. Por consiguiente, los resultados de la tabla 5-5 y de la figura 5-18 sugieren lo siguiente.

Propiedades de la distribución de proporciones muestrales

- **Las proporciones muestrales tienden a coincidir con el valor de la proporción poblacional.**
- **En ciertas condiciones, la distribución de proporciones muestrales se aproxima a una distribución normal.**

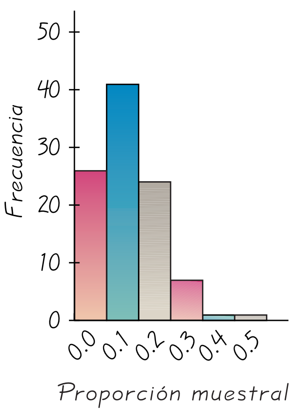


FIGURA 5-17 **100 proporciones muestrales con $n = 10$ en cada muestra**

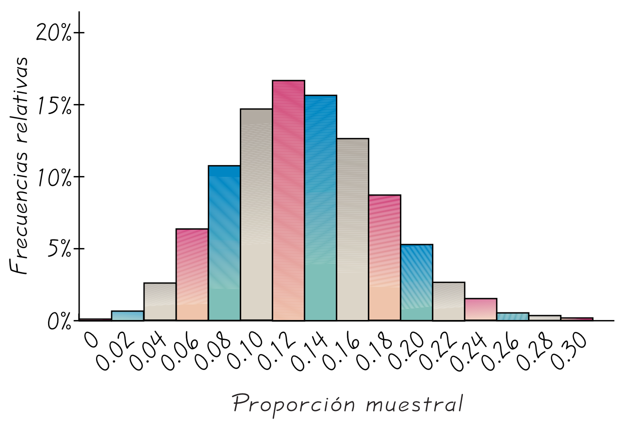


FIGURA 5-18 **10,000 proporciones muestrales con $n = 50$ en cada muestra**

¿Cuáles estadísticos son buenos estimadores de parámetros?

En el capítulo 6 estudiaremos métodos formales para el uso de estadísticos muestrales que nos permitirán hacer estimaciones de los valores de parámetros de población. Algunos estadísticos funcionan mucho mejor que otros, por lo cual es posible juzgar su valor examinando sus distribuciones muestrales, como en el siguiente ejemplo.

EJEMPLO Distribuciones muestrales Una *población* se compone de los valores 1, 2, 5. Si seleccionamos aleatoriamente muestras de tamaño 2 con reemplazo, hay nueve distintas muestras posibles, que se listan en la tabla 5-2. Como las nueve muestras distintas son igualmente posibles, cada muestra tiene una probabilidad de $1/9$.

a. Para cada muestra calcule la media, mediana, rango, varianza, desviación estándar y la proporción de valores muestrales impares. (Para cada estadístico, esto generará nueve valores que, cuando se asocien con nueve probabilidades de $1/9$ cada una, se combinarán para formar una *distribución muestral* del estadístico).

b. Para cada estadístico, calcule la media de los resultados del inciso *a*.

c. Compare las medias del inciso *b* con los parámetros de población correspondientes; después, determine si cada estadístico coincide con el valor del parámetro poblacional. Por ejemplo, las medias de muestra tienden a centrarse alrededor del valor de la media poblacional, que es $8/3 = 2.7$, de manera que la media de muestra coincide con el valor de la media poblacional.

SOLUCIÓN

a. Véase la tabla 5-2, que incluye los estadísticos individuales para cada muestra.

b. Las medias de los estadísticos muestrales aparecen casi al final de la tabla 5-2. La media de las medias de muestra es 2.7, la media de las medianas de muestra es 2.7, y así sucesivamente.

c. El renglón inferior de la tabla 5-2 se basa en una comparación de los parámetros poblacionales y los resultados de los estadísticos muestrales. Por ejemplo, la media poblacional de 1, 2, 5 es $\mu = 2.7$, y las medias de muestra "coinciden" con el valor de 2.7, ya que la media de las medias de muestra también es 2.7.

INTERPRETACIÓN Con base en los resultados de la tabla 5-2, observamos que cuando se utiliza un estadístico muestral para estimar un parámetro de población, algunos estadísticos son buenos porque coinciden con el parámetro poblacional y, por lo tanto, tienden a producir buenos resultados. Estadísticos como éstos se denominan *estimadores sin sesgo*. Otros estadísticos no son tan buenos (ya que son *estimadores sesgados*). He aquí un resumen.

- **Estadísticos que coinciden con los parámetros poblacionales:** media, varianza, proporción
- **Estadísticos que no coinciden con los parámetros poblacionales:** mediana, rango, desviación estándar

continúa

Aunque la desviación estándar muestral no coincide con la desviación estándar poblacional, el sesgo es relativamente pequeño en muestras grandes, de manera que con frecuencia utiliza s para estimar σ. Por consiguiente, a medias, proporciones, varianza y desviaciones estándar se les considerará temas importantes en los siguientes capítulos, pero la mediana y el rango se utilizarán en pocas ocasiones.

El objetivo más importante de esta sección es introducir el concepto de distribución muestral de un estadístico. Considere el objetivo de tratar de calcular la temperatura corporal media de todos los adultos. Puesto que la población es muy grande, no es práctico medir la temperatura de cada adulto. En su lugar, obtenemos una muestra de temperaturas corporales y la utilizamos para estimar la media poblacional. El conjunto de datos 4 del Apéndice B incluye una muestra de 106 temperaturas corporales; la media de esta muestra es $\bar{x} = 98.20°F$. Las conclusiones que hagamos acerca de la temperatura media poblacional de todos los adultos requiere que comprendamos el comportamiento de la distribución muestral de todas las medias de muestra de este tipo. Aunque no es práctico obtener cada muestra posible y nos conformamos con una sola muestra, es posible obtener algunas conclusiones muy importantes y con significado acerca de la población de todas las temperaturas corporales. Uno de los objetivos principales de las secciones y los capítulos que siguen es aprender el uso eficaz de una muestra para sacar conclusiones acerca de una población. En la sección 5-5 tendremos en cuenta más detalles acerca de la distribución muestral de medias de muestra, en tanto en la sección 5-6 estudiaremos más detalles acerca de la distribución muestral de las proporciones de muestra.

5-4 Destrezas y conceptos básicos

1. Encuesta de votantes Con base en una muestra aleatoria de $n = 400$ votantes, la división de noticias de la NBC predice que el candidato demócrata a la presidencia obtendrá el 49% de los votos, aunque en realidad obtiene el 51%. ¿Debemos concluir que la encuesta se realizó incorrectamente? ¿Por qué?

2. Distribución muestral de Harry Potter El conjunto de datos 14 del Apéndice B incluye una muestra de los niveles de lectura medidos para 12 páginas que se seleccionaron aleatoriamente del libro *Harry Potter y la piedra filosofal*, de J. K. Rowling. La media de los 12 valores del nivel de Flesch-Kincaid es 5.08. El valor de 5.08 forma parte de una distribución muestral. Describa esta distribución muestral.

3. Distribución muestral de temperaturas corporales El conjunto de datos 4 del Apéndice B incluye una muestra de 106 temperaturas corporales de adultos. Si construyésemos un histograma para describir la forma de la distribución de dicha muestra, ¿mostraría el histograma la forma de una *distribución muestral* de medias de muestra? ¿Por qué?

4. Distribución muestral de resultados de encuesta La organización Gallup realizó una encuesta a 1015 estudiantes que se seleccionaron al azar, desde jardín de niños hasta preparatoria, y encontró que el 10% acudía a escuelas privadas o religiosas.

a. ¿Será el resultado del 10% (o 0.10) un estadístico o un parámetro? Explique.

b. ¿Cuál es la distribución muestral sugerida por los datos?

c. ¿Tendría más confianza en los resultados si el tamaño de muestra hubiese sido de 2000 en lugar de 1025? ¿Por qué?

5. Centro telefónico La Nome Ice Company abrió únicamente durante tres días (adivine por qué). He aquí el número de llamadas que se recibieron durante cada uno de esos días: 10, 6, 5. Suponga que se seleccionan aleatoriamente muestras de tamaño 2, *con reemplazo*, de esta población de tres valores.

a. Liste las nueve muestras diferentes posibles y calcule la media de cada una de ellas.
b. Identifique la probabilidad de cada muestra y describa la distribución muestral de las medias de muestra. (*Sugerencia*: Véase la tabla 5-3).
c. Calcule la media de la distribución muestral.
d. ¿Será igual la media de la distribución muestral (del inciso *c*) a la media de la población de los tres valores que se listan? ¿Serán estas medias *siempre* iguales?

6. Telemercadeo A continuación se presenta el número de ventas por día de Kim Ryan, un cortés vendedor que trabajó durante cuatro días antes de que lo despidieran: 1, 11, 9, 3. Suponga que se seleccionan al azar muestras de tamaño 2, *con reemplazo*, de esta población de cuatro valores.

a. Liste las 16 diferentes muestras posibles y calcule la media de cada una de ellas.
b. Identifique la probabilidad de cada muestra y después describa la distribución muestral de medias de muestra. (*Sugerencia*: Véase la tabla 5-3).
c. Calcule la media de la distribución muestral.
d. ¿Es la media de la distribución muestral (del inciso *c*) igual a la media de la población de los cuatro valores que se listan? ¿Serán estas medias *siempre* iguales?

7. Estaturas de los Lakers de L.A. A continuación se presentan las estaturas (en pulgadas) de cinco jugadores de basquetbol estelares de los Lakers de Los Ángeles: 85, 79, 82, 73, 78. Suponga que se seleccionan al azar muestras de tamaño 2, *con reemplazo*, de la población de cinco estaturas.

a. Después de identificar las 25 distintas muestras posibles, calcule la media de cada una de ellas.
b. Describa la distribución muestral de medias. (*Sugerencia:* Véase la tabla 5-2).
c. Calcule la media de la distribución muestral.
d. ¿Es la media de la distribución muestral (del inciso *c*) igual a la media de la población de las cinco estaturas que se listan? ¿Serán estas medias *siempre* iguales?

8. Presidentes militares A continuación se lista la población de los cinco presidentes de Estados Unidos que tuvieron profesiones militares, junto con sus edades en el momento de tomar posesión: Eisenhower (62), Grant (46), Harrison (68), Taylor (64) y Washington (57). Suponga que se seleccionan al azar muestras de tamaño 2, *con reemplazo*, de la población de las cinco edades.

a. Después de identificar las 25 distintas muestras posibles, calcule la media de cada una de ellas.
b. Describa la distribución muestral de medias. (*Sugerencia:* Véase la tabla 5-2).
c. Calcule la media de la distribución muestral.
d. ¿Es la media de la distribución muestral (del inciso *c*) igual a la media de la población de las cinco edades? ¿Serán estas medias *siempre* iguales?

9. Genética Un experimento en genética incluye una población de moscas de la fruta consistente en un macho que llamaron Mike y tres hembras cuyos nombres son Anna, Bárbara y Chris. Suponga que se seleccionan dos moscas de la fruta al azar, *con reemplazo*.

a. Después de identificar las 16 diferentes muestras posibles, calcule la proporción de hembras en cada una de ellas.
b. Describa la distribución muestral de proporciones de hembras. (*Sugerencia*: Véase la tabla 5-2).
c. Calcule la media de la distribución muestral.
d. ¿Es la media de la distribución muestral (del inciso *c*) igual a la proporción poblacional de hembras? ¿Coincide *siempre* la media de la distribución muestral de proporciones con la proporción poblacional?

10. Control de calidad Después de construir una máquina nueva, se producen cinco faros prototipo para automóvil y se descubre que dos están defectuosos (D) y tres son aceptables (A). Suponga que se seleccionan dos faros al azar, *con reemplazo*, de esta población.
  a. Después de identificar las 25 distintas muestras posibles, calcule la proporción de defectos en cada una de ellas.
  b. Describa la distribución muestral de proporciones de defectos. (*Sugerencia*: Véase la tabla 5-2).
  c. Calcule la media de la distribución muestral.
  d. ¿Es la media de la distribución muestral (del inciso *c*) igual a la proporción poblacional de defectos? ¿Coincide *siempre* la media de la distribución muestral de proporciones con la proporción poblacional?

11. Senadoras Permitamos que una población se conforme con 10 senadoras demócratas y tres republicanas en el Congreso 107o de Estados Unidos.
  a. Desarrolle un procedimiento para seleccionar aleatoriamente (con reemplazo) una muestra de tamaño 5 de la población de 10 demócratas y tres republicanas; después, seleccione una muestra de éstas y liste los resultados.
  b. Calcule la proporción de demócratas en la muestra del inciso *a*.
  c. ¿Es la proporción del inciso *b* un estadístico o un parámetro?
  d. ¿Es la proporción muestral del inciso *b* igual a la proporción poblacional de demócratas? ¿Podrá cualquier muestra aleatoria de tamaño 5 resultar en una proporción muestral que iguale la proporción poblacional?
  e. Suponga que se listan todas las distintas muestras posibles de tamaño 5 y que se calcula la proporción muestral de cada una de ellas. ¿Qué se concluye acerca del valor de la media de estas proporciones muestrales?

12. Senadoras Permitamos que una población conste de los siguientes estados de residencia de las tres senadoras republicanas del Congreso 107o de Estados Unidos: Maine, Maine, Texas. Suponiendo que las muestras de tamaño 2 se seleccionan aleatoriamente de esta población, *sin* reemplazo, liste las distintas muestras posibles. Calcule la probabilidad de cada muestra. Además, para cada muestra calcule la proporción de senadoras de Maine. Por ejemplo, la muestra de “Maine y Texas” resulta en una proporción muestral de $1/2$ (porque una de las dos senadoras es de Maine). Calcule la media de la distribución muestral y verifique que sea igual a la proporción poblacional de senadoras de Maine.

5-4 Más allá de lo básico

13. A continuación se lista la población de los cinco presidentes de Estados Unidos que tuvieron profesiones militares, junto con sus edades en el momento de tomar posesión: Eisenhower (62), Grant (46), Harrison (68), Taylor (64) y Washington (57). Suponga que todas las muestras se seleccionan *sin reemplazo*.
  a. Después de listar todas las muestras posibles de tamaño $n = 2$, calcule la media y la desviación estándar de las medias de muestra.
  b. Después de listar todas las muestras posibles de tamaño $n = 3$, calcule la media y la desviación estándar de las medias de muestra.
  c. Después de listar todas las muestras posibles de tamaño $n = 4$, calcule la media y la desviación estándar de las medias de muestra.
  d. Cuando se realiza muestreo sin reemplazo, ¿tienden las medias de muestra a coincidir con el valor de la media poblacional?
  e. Con base en los resultados anteriores, ¿de qué manera se afecta la *variación* de la distribución muestral de medias de muestras al incrementar el tamaño de la muestra?

14. Desviación media absoluta La población de 1, 2, 5 se utilizó para elaborar la tabla 5-2. Identifique la distribución muestral de la desviación media absoluta (que se definió en la sección 2-5); después, determine si la desviación media absoluta de una muestra es un buen estadístico para estimar la desviación media absoluta de la población.

15. La mediana como estimador En la tabla 5-2 la distribución muestral de las medianas tiene una media de 2.7. Puesto que la media poblacional también es de 2.7, parecería que la mediana es un buen estadístico para estimar el valor de la media poblacional. Utilice los valores poblacionales 1, 2, 5, y calcule las 27 muestras de tamaño $n = 3$ que es posible seleccionar sin reemplazo; después, calcule la mediana y la media de cada una de las 27 muestras. Una vez con estos resultados, calcule la media de la distribución muestral de la mediana y la media de la distribución muestral de la media. Compare los resultados con la media poblacional de 2.7. ¿Qué concluye?

5-5 Teorema del límite central

Esta sección es *sumamente* importante porque presenta el teorema del límite central, que establece los fundamentos para estimar parámetros poblacionales y pruebas de hipótesis (temas que se estudian con profundidad en los siguientes capítulos). Mientras estudia esta sección, trate de evitar la confusión que causa el hecho de que el teorema del límite central implica dos distribuciones diferentes: la distribución de la población original y la distribución de las medias de muestra.

Los siguientes términos y conceptos clave se presentaron en secciones anteriores:

- Una *variable aleatoria* es una variable que tiene un solo valor numérico, determinado por el azar, para cada resultado de un procedimiento (sección 4-2).
- Una *distribución de probabilidad* es una gráfica, tabla o fórmula que da la probabilidad para cada valor de una variable aleatoria (sección 4-2).
- La *distribución muestral de la media* es la distribución de probabilidad de medias de muestra, donde todas las muestras tienen el mismo tamaño n (sección 5-4).

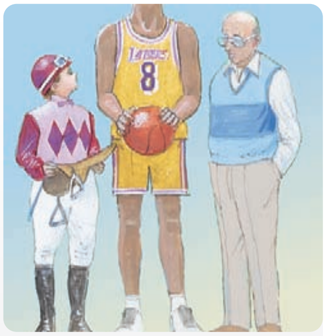
Véase el siguiente ejemplo, a modo de ilustración de estos conceptos abstractos.

EJEMPLO Dígitos aleatorios Considere la población de los dígitos 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, los cuales se seleccionan aleatoriamente con reemplazo.

a. *Variable aleatoria:* Si realizamos ensayos que consisten en la selección aleatoria de un solo dígito, y si representamos el valor del dígito que seleccionamos con x, entonces x es una variable aleatoria (porque su valor depende del azar).

b. *Distribución de probabilidad:* Suponiendo que los dígitos se seleccionan aleatoriamente, la probabilidad de cada dígito es $1/10$, que puede expresarse con la fórmula $P(x) = 1/10$. Ésta es una distribución de probabilidad (ya que describe la probabilidad de cada valor de la variable aleatoria x).

c. *Distribución muestral:* Ahora suponga que seleccionamos aleatoriamente todas las distintas muestras posibles de tamaño $n = 4$. (Recuerde que estamos haciendo muestreo con reemplazo, de modo que cualquier muestra particular tendría el mismo dígito más de una vez). En cada muestra calculamos la media de muestra $\bar{x}$ (que en sí misma es una variable aleatoria porque su valor depende del azar). La distribución de probabilidad de las medias muestrales es una distribución muestral.



El teorema del límite central borroso

En *The Cartoon Guide to Statistics*, de Gonick y Smith, los autores describen el teorema del límite central borroso de la siguiente manera: "Los datos que se ven influidos por efectos aleatorios muy pequeños y sin relación entre sí, se distribuyen aproximadamente de manera normal. Esto explica por qué la normalidad está en todos lados: fluctuaciones del mercado de las acciones, pesos de estudiantes, promedios anuales de temperatura, calificaciones del SAT: todos son el resultado de muchos efectos diferentes". La estatura de las personas, por ejemplo, es el resultado de factores hereditarios, factores ambientales, nutrición, cuidado de la salud, región geográfica y otras influencias que, cuando se combinan, producen valores distribuidos de forma normal.

El inciso *c* del ejemplo ilustra una distribución muestral específica de medias de muestra. En la sección 5-4 observamos que la media de medias de muestra es igual a la media de la población y que si, el tamaño de la muestra aumenta, las medias de muestra correspondientes tienden a variar menos. El *teorema del límite central* nos indica que si el tamaño de la muestra es lo suficientemente grande, la distribución de medias de muestra puede aproximarse a una *distribución normal*, aun si la población original no se distribuye normalmente. Aunque hablamos de un "teorema", no incluimos pruebas rigurosas; en su lugar, nos enfocamos en los *conceptos* y la forma de aplicarlos. A continuación se presentan los puntos clave, que conforman un fundamento importante para los siguientes capítulos.

El teorema del límite central y la distribución muestral de $\bar{x}$

Puesto que:

1. La variable aleatoria x tiene una distribución (que puede o no ser normal) con media μ y desviación estándar σ.
2. Todas las muestras aleatorias del mismo tamaño n se seleccionan de la población. (Las muestras se seleccionan de modo que todas las muestras posibles de tamaño n tengan la misma posibilidad de seleccionarse).

Conclusiones:

1. La distribución de las medias de muestra $\bar{x}$ se aproximará a una distribución *normal*, conforme el tamaño de la muestra aumente.
2. La media de todas las medias de muestra es la media poblacional μ. (Es decir, la distribución normal de la conclusión 1 tiene una media μ).
3. La desviación estándar de todas las medias de muestra es $\sigma/\sqrt{n}$. (Es decir, la distribución normal de la conclusión 1 tiene una desviación estándar $\sigma/\sqrt{n}$).

Reglas prácticas de uso común

1. Si la población original no se distribuye normalmente, la siguiente es una guía común: para muestras de tamaño n mayores que 30, la distribución de las medias de muestra puede aproximarse razonablemente bien a una distribución normal. (Hay excepciones, como las poblaciones con distribuciones muy diferentes a la normal, que requieren tamaños de muestra mucho más grandes que 30, aunque éstas son relativamente raras). La aproximación mejora conforme el tamaño muestral n se incrementa.
2. Si la población original se distribuye normalmente, entonces las medias de muestra se distribuirán normalmente para *cualquier* tamaño de muestra n (no sólo los valores de n mayores que 30).

El teorema del límite central implica dos distribuciones diferentes: la distribución de la población original y la distribución de las medias de muestra. Igual que en capítulos anteriores, utilizamos los símbolos μ y σ para denotar la media y la desviación estándar de la población original, pero ahora necesitamos nuevas notaciones para la media y la desviación estándar de la distribución de medias de muestra.

Notación para la distribución muestral de $\bar{x}$

Si se seleccionan todas las muestras aleatorias de tamaño n de una población con media μ y desviación estándar σ, la media de las medias de muestra se denota con $\mu_{\bar{x}}$, de modo que

$$\mu_{\bar{x}} = \mu$$

También la desviación estándar de las medias de muestra se denota con $\sigma_{\bar{x}}$, de manera que

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$\sigma_{\bar{x}}$ suele denominarse el **error estándar de la media.**

EJEMPLO Dígitos aleatorios Nuevamente considere la población de dígitos 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, que se seleccionan aleatoriamente con reemplazo. Suponga que seleccionamos al azar muestras de tamaño $n = 4$. En la población original de dígitos, todos los valores son igualmente probables. Con base en las “Reglas prácticas de uso común” (que se incluyen en el recuadro del teorema del límite central), *no podemos* concluir que las medias de muestra están distribuidas normalmente, ya que la población original no tiene una distribución normal y el tamaño de muestra 4 no es mayor que 30. Sin embargo, exploraremos la distribución muestral para ver qué aprendemos de ella.

La tabla 5-6 se elaboró registrando los cuatro últimos dígitos de los números de seguridad social de cada uno de 50 estudiantes. Los últimos cuatro dígitos de los números del seguro social son aleatorios, a diferencia de los dígitos iniciales, que se utilizan para codificar información particular. Si combinamos los cuatro dígitos de cada estudiante en un conjunto grande de 200 números, obtendremos una media $\bar{x} = 4.5$, una desviación estándar $s = 2.8$, y una distribución como la gráfica que se presenta en la tabla 5-19. Ahora note qué ocurre cuando calculamos las 50 medias de las muestras, como se observa en la tabla 5-6. (Por ejemplo, el primer estudiante tiene los dígitos 1, 8, 6 y 4, en tanto que la media de los cuatro dígitos es 4.75). Aun cuando el conjunto original de datos *no* tiene una distribución normal, las medias de muestra presentan una distribución aproximadamente *normal*. Esto tal vez resulte confuso, por lo que conviene detenerse aquí y estudiar el párrafo hasta que su idea central quede clara: el conjunto original de los 200 números individuales *no* tiene una distribución normal (porque los dígitos 0-9 ocurren con frecuencias aproximadamente iguales), pero las 50 medias de muestra *sí* tienen una distribución aproximadamente normal. (Una de las “Reglas prácticas de uso común” establece que las muestras con $n > 30$ pueden aproximarse a una distribución normal, pero las muestras pequeñas, como $n = 4$ de este ejemplo, en ocasiones tienen distribuciones aproximadamente normales). El hecho de que al hacer un muestreo de una distribución lleguemos a crear una distribución de medias de muestra que sea normal, o al menos aproximadamente normal, es un fenómeno verdaderamente fascinante e intrigante de la estadística.

Tabla 5-6

SSN dígitos				$\bar{x}$
1	8	6	4	4.75
5	3	3	6	4.25
9	8	8	8	8.25
5	1	2	5	3.25
9	3	3	5	5.00
4	2	6	2	3.50
7	7	1	6	5.25
9	1	5	4	4.75
5	3	3	9	5.00
7	8	4	1	5.00
0	5	6	1	3.00
9	8	2	2	5.25
6	1	5	7	4.75
8	1	3	0	3.00
5	9	6	9	7.25
6	2	3	4	3.75
7	4	0	7	4.50
5	7	5	6	5.75
4	1	5	7	4.25
1	2	0	6	2.25
4	0	2	8	3.50
3	1	2	5	2.75
0	3	4	0	1.75
1	5	1	0	1.75
9	7	4	0	5.00
7	3	1	1	3.00
9	1	1	3	3.50
8	6	5	9	7.00
5	6	4	1	4.00
9	3	9	5	6.50
6	0	7	3	4.00
8	2	9	6	6.25
0	2	8	6	4.00
2	0	9	7	4.50
5	8	9	0	5.50
6	5	4	9	6.00
4	8	7	6	6.25
7	1	2	0	2.50
2	9	5	0	4.00
8	3	2	2	3.75
2	7	1	6	4.00
6	7	7	1	5.25
2	3	3	9	4.25
2	4	7	5	4.50
5	4	3	7	4.75
0	4	3	8	3.75
2	5	8	6	5.25
7	1	3	4	3.75
8	3	7	0	4.50
5	6	6	7	6.00

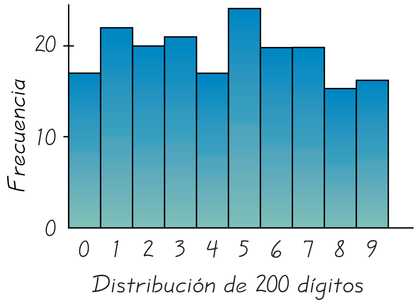


FIGURA 5-19 Distribución de 200 dígitos de los números del seguro social

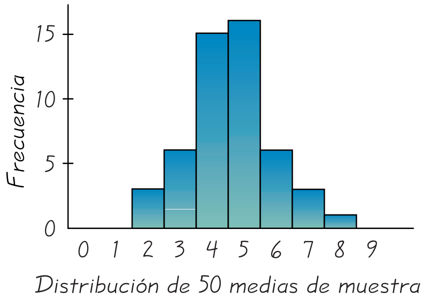


FIGURA 5-20 Distribución de 50 medias de muestra

En la figura 5-20 vemos que la distribución de las medias de muestra del ejemplo anterior es aproximadamente normal, a pesar de que la población original no tiene una distribución normal y de que el tamaño de $n = 4$ de las muestras individuales no excede de 30. Si examina con atención la figura 5-20 verá que no se trata de una distribución normal exacta, pero se acercaría a una distribución normal exacta conforme el tamaño de la muestra se incrementa más allá de 4.

Conforme aumenta el tamaño de la muestra, la distribución muestral de medias de muestra se aproxima a una distribución *normal*.

Aplicación del teorema del límite central

Es posible resolver muchos problemas prácticos importantes con el teorema del límite central. Cuando trabaje en este tipo de problemas, recuerde que si el tamaño de la muestra es mayor que 30, o si la población original se distribuye normalmente, debe tratar la distribución de medias de muestra como si fuera una distribución normal con media μ y desviación estándar $\sigma/\sqrt{n}$.

En el siguiente ejemplo, el inciso *a*) incluye un valor *individual*, pero el inciso *b* incluye la media de una *muestra* de 36 mujeres, por lo cual usaremos el teorema del límite central al trabajar con la variable aleatoria $\bar{x}$. Estudie este ejemplo con atención para comprender la diferencia significativa entre los procedimientos que se utilizaron en los incisos *a* y *b*. Observe cómo este ejemplo ilustra el siguiente procedimiento de trabajo:

- **Cuando trabaje con un valor *individual* de una población que se distribuye normalmente, utilice los métodos de la sección 5-3. Use $z = \dfrac{x - \mu}{\sigma}$.**
- **Cuando trabaje con una media de alguna *muestra* (o grupo), asegúrese de utilizar el valor de $\sigma/\sqrt{n}$ para la desviación estándar de las medias de muestra. Use $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$.**

PROBLEMA 5 DEL CAPÍTULO

EJEMPLO Seguridad del teleférico En el problema del capítulo señalamos que un teleférico en Vail, Colorado, lleva a los esquiadores a la cima de la montaña. Hay una placa que indica que su capacidad máxima es de 12 personas o 2004 libras. Dicha capacidad se excedería si 12 personas tienen pesos con una media mayor que $2004/12 = 167$ libras.

Puesto que los hombres suelen pesar más que las mujeres, el "peor de los casos" implicaría a 12 pasajeros hombres. Los pesos de los hombres se distribuyen normalmente, con una media de 172 libras y una desviación estándar de 29 libras (según datos del National Health Survey).

a. Calcule la probabilidad de que, al seleccionar aleatoriamente a un hombre, su peso sea mayor de 167 libras.

b. Calcule la probabilidad de que 12 hombres que se seleccionaron al azar tengan una media mayor de 167 libras (de manera que su peso total sea mayor que la máxima capacidad del teleférico de 2004 libras).

SOLUCIÓN

a. *Aproximación: Utilice los métodos presentados en la sección 5-3* (ya que estamos trabajando con un valor *individual* de una población distribuida normalmente). Buscamos el área de la región que se sombreó de la figura 5-21a. Antes de emplear la tabla A-2, convertimos el peso de 167 a su puntuación z correspondiente:

$$z = \frac{x - \mu}{\sigma} = \frac{167 - 172}{29} = -0.17$$

Ahora nos remitimos a la tabla A-2, usando $z = -0.17$; encontramos que el área acumulativa a la izquierda de 167 libras es 0.4325. La región que se sombreó es, por tanto, $1 - 0.4325 = 0.5675$. La probabilidad de que un hombre que se selecciona aleatoriamente pese más de 167 libras es de 0.5675.

b. *Aproximación: Utilice el teorema del límite central* (ya que estamos trabajando con la *media de una muestra* de 12 hombres, no de un solo hombre). Aun cuando el tamaño de la muestra no es mayor de 30, empleamos una distribución normal por la siguiente razón: la población original de hombres tiene una distribución normal, de manera que las muestras de cualquier tamaño producirán medias distribuidas normalmente. Puesto que ahora estamos trabajando con una distribución de medias de muestra, debemos emplear los parámetros $\mu_{\bar{x}}$ y $\sigma_{\bar{x}}$, que se evalúan de la siguiente manera:

$$\mu_{\bar{x}} = \mu = 172$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{29}{\sqrt{12}} = 8.37158$$

continúa

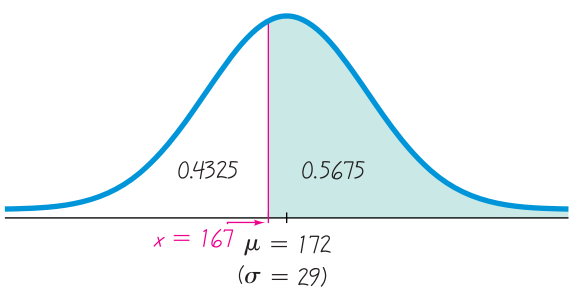


FIGURA 5-21(a) **Distribución de pesos individuales de hombres**

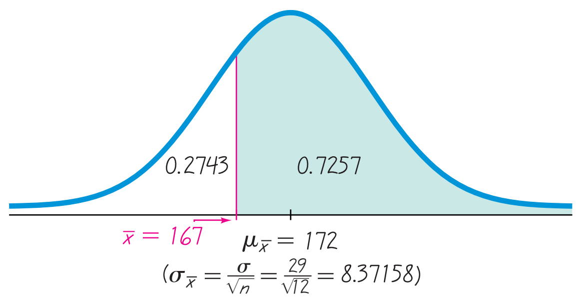


FIGURA 5-21(b) **Medias de muestras de los pesos de 12 hombres**

El siguiente es un punto importante: debemos usar la desviación estándar que se calculó de 8.37158, no la desviación estándar original de 29 (porque estamos trabajando con la distribución de medias de muestra, con una desviación estándar de 8.37158 y no con la distribución de pesos individuales, cuya desviación estándar es de 29). Deseamos encontrar el área que se sombreó de la figura 5-21b. En la tabla A-2 obtenemos la puntuación z relevante, que se calcula de la siguiente manera:

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{167 - 172}{\frac{29}{\sqrt{12}}} = \frac{-5}{8.37158} = -0.60$$

Si nos remitimos a la tabla A-2, encontramos que $z = -0.60$ corresponde a un área izquierda acumulativa de 0.2743, por lo que la región que se sombreó es $1 - 0.2743 = 0.7257$. La probabilidad de que los 12 hombres tengan un peso medio mayor que 167 libras es de 0.7257.

INTERPRETACIÓN Hay una probabilidad de 0.5675 de que un hombre pese más de 167 libras y otra de 0.7257 de que 12 hombres tengan un peso medio mayor que 167 libras. Puesto que la capacidad máxima del teleférico es de 2004 libras, es probable (con una probabilidad de 0.7257) que se sobrecargue si se llena con 12 hombres que se seleccionaron al azar. Sin embargo, la seguridad de los pasajeros no es tan mala, por factores como: **1.** es probable que los hombres esquiadores tengan un peso medio menor que la media de 172 libras de la población general de hombres; **2.** es probable que también suban mujeres esquiadoras, y tienden a pesar menos que los hombres; **3.** a pesar de que la capacidad máxima que se señaló es de 2004 libras, el teleférico se diseñó para operar con seguridad con pesos muy por encima de la carga conservadora de 2004 libras. No obstante, los operadores del teleférico harían bien en evitar una carga de 12 hombres, especialmente si parecen ser muy pesados. Los cálculos utilizados aquí son exactamente los mismos que emplean los ingenieros cuando diseñan teleféricos, elevadores, escaleras eléctricas, aviones y otros aparatos que cargan personas.

Interpretación de resultados

El siguiente ejemplo ilustra otra aplicación del teorema del límite central, pero examine con atención la conclusión a la que se llega. Este ejemplo muestra el tipo de pensamiento que es fundamental para el importante procedimiento de prueba de hipótesis (que se estudia en el capítulo 7). Este ejemplo ilustra la regla del suceso poco común de la estadística inferencial, que se presentó inicialmente en la sección 3-1.

Regla del suceso poco común

Si, bajo cierto supuesto, la probabilidad de un suceso particular que se observa es excepcionalmente pequeña, concluimos que el supuesto probablemente no es correcto.

EJEMPLO Temperaturas corporales Suponga que la población de temperaturas corporales humanas tiene una media de 98.6°F, como suele creerse. También que la desviación estándar de la población es 0.62°F (de acuerdo con datos de investigadores de la Universidad de Maryland). Si se selecciona al azar una muestra de tamaño $n = 106$, calcule la probabilidad de obtener una media de 98.2°F o menor. (En realidad, se obtuvo el valor de 98.2°F; observe las temperaturas de medianoche del día 2 en el conjunto de datos 4 del Apéndice B).

SOLUCIÓN No se nos dio la distribución de la población pero, porque el tamaño de la muestra $n = 106$ excede a 30, utilizamos el teorema del límite central y concluimos que la distribución de medias de muestra es normal, con estos parámetros:

$$\mu_{\bar{x}} = \mu = 98.6 \qquad \text{(por supuesto)}$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.62}{\sqrt{106}} = 0.0602197$$

La figura 5-22 muestra el área que se sombreó (observe la pequeña cola izquierda de la gráfica), correspondiente a la probabilidad que buscamos. Con los parámetros que se aplican a la distribución de la figura 5-22, hallaremos el área que se sombreó utilizando los mismos procedimientos desarrollados en la sección 5-3. En la tabla A-2 primero encontramos la puntuación z:

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{98.20 - 98.6}{0.0602197} = -6.64$$

continúa

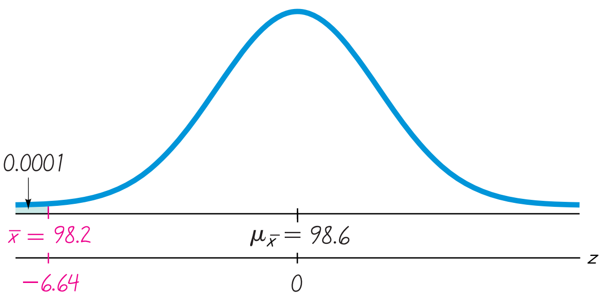


FIGURA 5-22 Distribución de temperaturas corporales medias, para muestras de tamaño $n = 106$

Si nos remitimos a la tabla A-2, encontraremos que $z = -6.64$ no aparece, pero para los valores de z que están por debajo de -3.49 utilizamos un área de 0.0001 para el área izquierda acumulativa hasta $z = -3.49$. Por lo tanto, concluimos que la región que se sombreó de la figura 5.22 es 0.0001. (Las tablas más precisas de los resultados de la calculadora TI-83 Plus indican que el área de la región que se sombreó es cercana a 0.00000000002, pero incluso tales resultados son sólo aproximaciones. Con seguridad reportaríamos que la probabilidad es muy baja, menor que 0.001).

INTERPRETACIÓN Los resultados demuestran que si la media de nuestra temperatura corporal es en realidad 98.6°F, entonces hay una probabilidad sumamente baja de obtener una media de muestra de 98.2°F o menor, cuando se seleccionan 106 sujetos aleatoriamente. Los investigadores de la Universidad de Maryland obtuvieron una media muestral como ésta, y existen dos explicaciones posibles: o la media de la población es en realidad de 98.6°F y su muestra representa un suceso aleatorio extremadamente poco común, o la media poblacional es menor que 98.6°F y su muestra es típica. Como la probabilidad es tan baja, parece más razonable concluir que la media poblacional es menor que 98.6°F. Éste es el tipo de razonamiento que se usa en la *prueba de hipótesis*, que se estudiará en el capítulo 7. Por ahora, habrá que enfocarnos en el uso del teorema del límite central para calcular la probabilidad de 0.0001, pero debemos tomar en cuenta que dicho teorema se utilizará posteriormente para explicar algunos conceptos muy importantes en estadística.

Corrección para una población finita

Al aplicar el teorema del límite central, el uso de $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ supone que la población tiene un número infinito de miembros. Cuando hacemos un muestreo con reemplazo (es decir, que se regresa cada elemento que se eligió antes de hacer la siguiente selección), la población es efectivamente infinita. Aunque muchas aplicaciones realistas implican un muestreo sin reemplazo, tales muestras sucesivas dependen de resultados previos. En la fabricación, los inspectores de control de calidad suelen muestrear elementos de una racha finita de producción, sin reemplazarlos. Para una población finita como ésta tal vez necesitemos ajustar $\sigma_{\bar{x}}$. La siguiente es una regla práctica:

Cuando realice un muestreo sin reemplazo y el tamaño de muestra n sea mayor que el 5% de la población finita de tamaño N (es decir, $n > 0.05N$), ajuste la desviación estándar de medias de muestra $\sigma_{\bar{x}}$, multiplicándola por el *factor de corrección de población finita:*

$$\sqrt{\frac{N - n}{N - 1}}$$

Con excepción de los ejercicios 21 y 22, los ejemplos y los ejercicios de esta sección suponen que el factor de corrección de población finita *no* se aplica, porque estamos tomando una muestra con reemplazo, porque la población es infinita o porque el tamaño de muestra no excede el 5% del tamaño de la población.

El teorema del límite central es muy importante porque nos permite usar los métodos básicos de la distribución normal en una amplia variedad de circunstancias. Por ejemplo, en el capítulo 6 aplicaremos el teorema cuando utilicemos datos muestrales para estimar medias de poblaciones. En el capítulo 7 lo aplicaremos cuando usemos datos muestrales para probar aseveraciones hechas sobre medias poblacionales. Dichas aplicaciones para estimar parámetros de población y probar aseveraciones, constituyen usos sumamente importantes de la estadística, en tanto que el teorema del límite central los hace posibles.

5-5 Destrezas y conceptos básicos

Uso del teorema del límite central. *En los ejercicios 1 a 6, suponga que los pesos de los hombres se distribuyen de manera normal, con una media dada por $\mu = 172$ libras y una desviación estándar dada por $\sigma = 29$ libras (según datos del National Health Survey).*

1. **a.** Si se selecciona un hombre al azar, calcule la probabilidad de que pese menos de 167 libras.
b. Si se seleccionan 36 hombres al azar, calcule la probabilidad de que tengan un peso medio menor de 167 libras.

2. **a.** Si se selecciona un hombre al azar, calcule la probabilidad de que su peso sea mayor de 180 libras.
b. Si se seleccionan 100 hombres al azar, calcule la probabilidad de que tengan un peso medio mayor de 180 libras.

3. **a.** Si se selecciona un hombre al azar, calcule la probabilidad de que su peso se ubique entre 170 y 175 libras.
b. Si se seleccionan 64 hombres al azar, calcule la probabilidad de que tengan un peso medio entre 170 y 175 libras.

4. **a.** Si se selecciona un hombre al azar, calcule la probabilidad de que pese entre 100 y 165 libras.
b. Si se seleccionan 81 hombres al azar, calcule la probabilidad de que tengan un peso medio entre 100 y 165 libras.

5. **a.** Si se seleccionan 25 hombres al azar, calcule la probabilidad de que tengan un peso medio mayor de 160 libras.
b. ¿Por qué puede usarse el teorema del límite central en el inciso *a*, a pesar de que el tamaño de muestra no excede a 30?

6. **a.** Si se seleccionan cuatro hombres al azar, calcule la probabilidad de que tengan un peso medio entre 160 y 180 libras.
b. ¿Por qué puede usarse el teorema del límite central en el inciso *a*, a pesar de que el tamaño de muestra no excede a 30?

7. Rediseño de asientos de expulsión En el problema del capítulo se señaló que los ingenieros estaban rediseñando los asientos de expulsión de aviones de combate para que se ajustaran mejor a las mujeres. Antes de que las mujeres se convirtieran en pilotos de aviones de combate, los asientos expulsores ACES-II se diseñaron para hombres que pesaran entre 140 libras y 211 libras. La población de mujeres tiene pesos distribuidos normalmente, con una media de 143 libras y una desviación estándar de 29 libras (de acuerdo con datos del National Health Survey).
a. Si se selecciona una mujer al azar, calcule la probabilidad de que pese entre 140 libras y 211 libras.
b. Si se seleccionan 36 mujeres diferentes al azar, calcule la probabilidad de que su peso medio se ubique entre 140 y 211 libras.
c. Al rediseñar los asientos expulsores de aviones de combate, para que se ajusten mejor a las mujeres, ¿qué probabilidad es más importante: el resultado del inciso *a* o el resultado del inciso *b*? ¿Por qué?

8. Diseño de cascos para motociclistas Los ingenieros deben tomar en cuenta la anchura de las cabezas de los hombres cuando diseñan cascos para motociclistas. Las anchuras de las cabezas de los hombres se distribuyen normalmente, con una media de 6.0 pulgadas y una desviación estándar de 1.0 pulgada (según datos de una encuesta antropométrica de Gordon, Churchill *et al*.).

a. Si se selecciona un hombre al azar, calcule la probabilidad de que el ancho de su cabeza sea menor que 6.2 pulgadas.

b. La compañía Safeguard Helmet planea una racha de producción inicial de 100 cascos. Calcule la probabilidad de que 100 hombres, que se seleccionaron al azar, tengan una anchura media de cabeza menor que 6.2 pulgadas.

c. El gerente de producción observa los resultados del inciso *b* y piensa que todos los cascos deben hacerse para hombres con anchuras de cabeza menores de 6.2 pulgadas, porque se ajustarían a casi todos los hombres. ¿Por qué es incorrecto este razonamiento?

9. Diseño de montaña rusa El Rock'n'Roller Coaster de los estudios MGM de Disney, en Orlando, tiene dos asientos en cada fila. Al diseñar esa montaña rusa, debía determinarse la anchura total de los dos asientos de cada fila. En el "peor de los casos", ambos asientos los ocupan hombres. Los hombres tienen anchuras de cadera distribuidas normalmente, con una media de 14.4 pulgadas y una desviación estándar de 1.0 pulgadas (según datos de una encuesta antropométrica de Gordon, Churchill *et al*.). Suponga que se seleccionan dos hombres al azar.

a. Calcule la probabilidad de que su anchura media de cadera sea mayor que 16.0 pulgadas.

b. Si cada fila de dos asientos se diseña para ajustarse a dos hombres, sólo si éstos tienen una anchura media de cadera de 16.0 pulgadas o menos, ¿muchos individuos serían incapaces de ajustarse? ¿Parece aceptable este diseño?

10. Generador uniforme de números aleatorios El generador de números aleatorios de la calculadora TI-83 Plus, así como el de muchas otras calculadoras y computadoras producen números de una distribución uniforme de valores entre 0 y 1, con una media de 0.500 y una desviación estándar de 0.289. Si se generan 100 números aleatorios, calcule la probabilidad de que su media sea mayor que 0.57. ¿Sería poco común generar 100 de estos números y obtener una media mayor que 0.57? ¿Por qué?

11. Cantidades de Coca Cola Suponga que latas de Coca Cola se llenan de tal manera que las cantidades reales tienen una media de 12.00 onzas y una desviación estándar de 0.11 onzas.

a. Calcule la probabilidad de que una muestra de 36 latas tenga una cantidad media de al menos 12.19 onzas, como en el conjunto de datos 17 del Apéndice B.

b. Con base en el resultado del inciso *a*, ¿será razonable creer que las latas en realidad contienen una media de 12.00 onzas? Si la media no es 12.00 onzas, ¿se está engañando a los consumidores?

12. Puntuaciones de CI Para formar parte de la organización Mensa se requiere una puntuación de CI por arriba de 131.5. Nueve candidatos toman pruebas de CI y el resumen de sus resultados indica que su puntuación media de CI es 133. (Las puntuaciones de CI se distribuyen de manera normal, con una media de 100 y una desviación estándar de 15).

a. Si se selecciona una persona al azar, calcule la probabilidad de elegir a alguien con una puntuación de CI de al menos 133.

b. Si se seleccionan nueve personas al azar, calcule la probabilidad de que su puntuación media de CI sea de al menos 133.

c. Aunque el resumen de los resultados está disponible, se perdieron las puntuaciones individuales de CI. ¿Se puede concluir que los nueve candidatos tienen puntuaciones de CI mayores de 133, de manera que todos son elegibles para formar parte de Mensa?

13. Tiempo medio de reemplazos El gerente de la tienda Portland Electronics se preocupa porque sus distribuidores le están entregando televisores con una calidad menor al promedio. Su investigación revela que los tiempos de reemplazo de televisores tienen una media de 8.2 años y una desviación estándar de 1.1 años (según datos de “Getting Things Fixed”, *Consumer Reports*). Entonces, selecciona al azar 50 televisores que se vendieron en el pasado y encuentra que el tiempo de reemplazo es de 7.8 años.

a. Suponiendo que el tiempo de reemplazo de televisores tiene una media de 8.2 años y una desviación estándar de 1.1 años, calcule la probabilidad de que 50 televisores que se seleccionaron aleatoriamente tengan un tiempo medio de reemplazo de 7.8 años o menos.

b. Con base en el resultado del inciso *a*, ¿parecería que la tienda Portland Electronics vende televisores con una calidad menor al promedio?

14. Presión sanguínea La presión sistólica (en mm Hg) de mujeres entre 18 y 24 años se distribuye normalmente, con una media de 114.8 y una desviación estándar de 13.1 (de acuerdo con datos del National Health Survey). La hipertensión suele definirse como una presión sistólica que rebasa 140.

a. Si se selecciona al azar a una mujer de entre 18 y 24 años, calcule la probabilidad de que su presión sistólica sea mayor que 140.

b. Si se seleccionan al azar cuatro mujeres del mismo rango de edad, calcule la probabilidad de que su presión sistólica media sea mayor que 140.

c. Puesto que el inciso *b* incluye un tamaño de muestra no mayor que 30, ¿por qué se puede utilizar el teorema del límite central?

d. Si un médico recibe un reporte que afirma que cuatro mujeres tienen una presión sistólica media menor que 140, ¿concluiría que ninguna de las mujeres es hipertensa (con una presión sanguínea mayor que 140)?

15. Reducción de nicotina en cigarrillos Las cantidades de nicotina en los cigarrillos Dytusoon tienen una media de 0.941 g y una desviación estándar de 0.313 g (con base en el conjunto de datos 5 del Apéndice B). La Huntington Tobacco Company, que produce los cigarrillos Dytusoon, afirma que redujo la cantidad de nicotina. La evidencia consiste en una muestra de 40 cigarrillos con una cantidad media de nicotina de 0.882 g.

a. Suponiendo que la media y la desviación estándar dadas no cambiaron, calcule la probabilidad de seleccionar al azar 40 cigarrillos con una cantidad media de nicotina de 0.882 g o menos.

b. Con base en el resultado del inciso *a*, ¿se valdrá afirmar que la cantidad de nicotina es menor? ¿Por qué?

16. Preparación para la prueba SAT Las calificaciones de hombres en la parte verbal de la prueba SAT-I se distribuyen normalmente, con una media de 509 y una desviación estándar de 112 (según datos del College Board). A hombres que se seleccionaron aleatoriamente se les da el Columbian Review Course, antes de tomar la prueba SAT. Suponga que el curso no tiene efecto alguno.

a. Si se selecciona a uno de los hombres al azar, calcule la probabilidad de que su calificación sea de al menos 590.

b. Si se selecciona a 16 de los hombres al azar, calcule la probabilidad de que su calificación media sea de al menos 590.

c. En el cálculo de la probabilidad del inciso *b*, ¿por qué puede usarse el teorema del límite central si el tamaño muestral no excede de 30?

d. Si la muestra aleatoria de 16 hombres arroja una calificación media de 590, ¿habría una fuerte evidencia que apoye la afirmación de que el curso es realmente eficaz? ¿Por qué?

17. Sobrecarga de un depósito de desperdicios La ciudad de Newport opera un depósito de basura que se sobrecarga si las descargas de desperdicios de sus 4872 hogares exceden una media de 27.88 libras en una semana. En muchas semanas se observa que las muestras de 4872 hogares tienen pesos que se distribuyen de manera normal, con una media de 27.44 libras y una desviación estándar de 12.46 libras (según datos del Garbage Project de la Universidad de Arizona). ¿Cuál es la proporción de semanas cuando el depósito de basura se sobrecarga? ¿Será un nivel aceptable o se deben tomar acciones para corregir un problema de un sistema que se sobrecargó?

18. Etiquetas de paquetes de M&M Los dulces M&M sencillos tienen un peso medio de 0.9147 g y una desviación estándar de 0.0369 g (con base en el conjunto de datos 19 del Apéndice B). Los dulces M&M que se utilizan en el conjunto de datos 19 provienen de un paquete que contenía 1498 dulces y la etiqueta del paquete establecía que su peso neto era de 1361 g. (Si cada paquete tiene 1498 dulces, el peso medio de los dulces debe exceder $1361/1498 = 0.9085$ g del contenido neto, para pesar al menos 1361 g).

a. Si se selecciona al azar un dulce M&M sencillo, calcule la probabilidad de que pese más de 0.9085 g.

b. Si se seleccionan al azar 1498 dulces M&M sencillos, calcule la probabilidad de que su peso medio sea de al menos de 0.9085 g.

c. Con estos resultados, ¿está Mars Company ofreciendo a los consumidores de M&M la cantidad que anuncia en la etiqueta?

19. Diseño de elevadores Los pesos de las mujeres se distribuyen normalmente, con una media de 143 libras y una desviación estándar de 29 libras, y los pesos de los hombres se distribuyen normalmente, con una media de 172 libras y una desviación estándar de 29 libras (datos del National Health Survey). Usted necesita diseñar un elevador para Wesport Shopping Center, el cual debe llevar a salvo a 16 personas. Suponiendo que “en el peor de los casos” se suban 16 pasajeros hombres, calcule el peso máximo total que se permite si deseamos un probabilidad de 0.975 de que este máximo no se rebase cuando se seleccione aleatoriamente a 16 hombres.

20. Diseño de asiento Usted necesita construir una banca que sentará a 18 jugadores universitarios de futbol americano y debe determinar primero la longitud de la banca. Los hombres tienen anchuras de cadera que se distribuyen normalmente, con una media de 14.4 pulgadas y una desviación estándar de 1.0 pulgadas.

a. ¿Cuál será la longitud mínima de la banca, si usted busca una probabilidad de 0.975 de que se ajuste a las anchuras de cadera de 18 hombres que se seleccionaron al azar?

b. ¿Por qué sería incorrecto utilizar realmente el resultado del inciso *a* como longitud de la banca?

5-5 Más allá de lo básico

21. Corrección para una población finita El club Boston Women necesita un elevador que se limite a ocho pasajeros. El club tiene 120 miembros mujeres con pesos que se aproximan a una distribución normal, con una media de 143 libras y una desviación estándar de 29 libras. (*Sugerencia:* Véase la explicación del factor de corrección para una población finita).

a. Si se seleccionan al azar ocho miembros diferentes, calcule la probabilidad de que su peso total no rebase la capacidad máxima de 1300 libras.

b. Si buscamos una probabilidad de 0.99 de que el elevador no se sobrecargue siempre que se seleccione aleatoriamente a ocho miembros como pasajeros, ¿cuál debe ser el peso máximo permitido?

22. Parámetros de población Una *población* se compone de los valores: 2, 3, 6, 8, 11, 18.

a. Calcule μ y σ.

b. Liste todas las muestras de tamaño $n = 2$ que es posible obtener con reemplazo.

c. Calcule la población de todos los valores de $\bar{x}$ al obtener la media de cada muestra del inciso *b*.

d. Calcule la media $\mu_{\bar{x}}$ y la desviación estándar $\sigma_{\bar{x}}$ para la población de medias de muestra obtenidas en el inciso *c*).

e. Verifique que

$$\mu_{\bar{x}} = \mu \qquad \text{y} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

23. Generador uniforme de números aleatorios En el ejercicio 10 se señaló que muchas calculadoras y computadoras tienen generadores de números aleatorios que producen números de una distribución uniforme de valores entre 0 y 1, con una media de 0.500 y una desviación estándar de 0.289. Si se generan 100 números aleatorios, calcule la probabilidad de que su media caiga entre 0.499 y 0.501. Si generásemos 100 de estos números y encontrásemos que la media cae entre 0.499 y 0.501, ¿concluiríamos que el resultado es " poco común", de tal manera que el generador de números aleatorios está defectuoso? ¿Por qué?

5-6 La distribución normal como aproximación de la distribución binomial

En lugar de "la distribución normal como aproximación de la distribución binomial", el encabezado apropiado para esta sección debería ser "uso de la distribución normal como aproximación de la distribución binomial", pero preferimos el primer título. El segundo refleja mejor el propósito de esta sección. Comencemos revisando las condiciones requeridas para una *distribución de probabilidad binomial*, que se analizan en la sección 4-3:

1. El procedimiento debe tener un *número fijo de ensayos*.
2. Los ensayos deben ser *independientes*.
3. Todos los resultados de cada ensayo deben estar clasificados en *dos categorías*.
4. Las probabilidades deben permanecer *constantes* para cada ensayo.

En la sección 4-3 presentamos tres métodos para calcular probabilidades binomiales: **1.** uso de la fórmula de probabilidad binomial, **2.** uso de la tabla A-1 y **3.** uso de programas de cómputo (tales como STATDISK, Minitab o Excel) o una calculadora TI-83 Plus. Sin embargo, en muchos casos ninguno de estos métodos es práctico, ya que los cálculos requieren demasiado tiempo y esfuerzo. Ahora presentamos un nuevo método que utiliza una distribución normal como aproximación de la distribución binomial. El siguiente recuadro resume el punto más importante de esta sección.

La distribución normal como aproximación de la distribución binomial

Si $np \geq 5$ y $nq \geq 5$, entonces la variable aleatoria binomial tiene una distribución de probabilidad que puede aproximarse con una distribución normal, donde la media y la desviación estándar están dadas por

$$\mu = np$$
$$\sigma = \sqrt{npq}$$

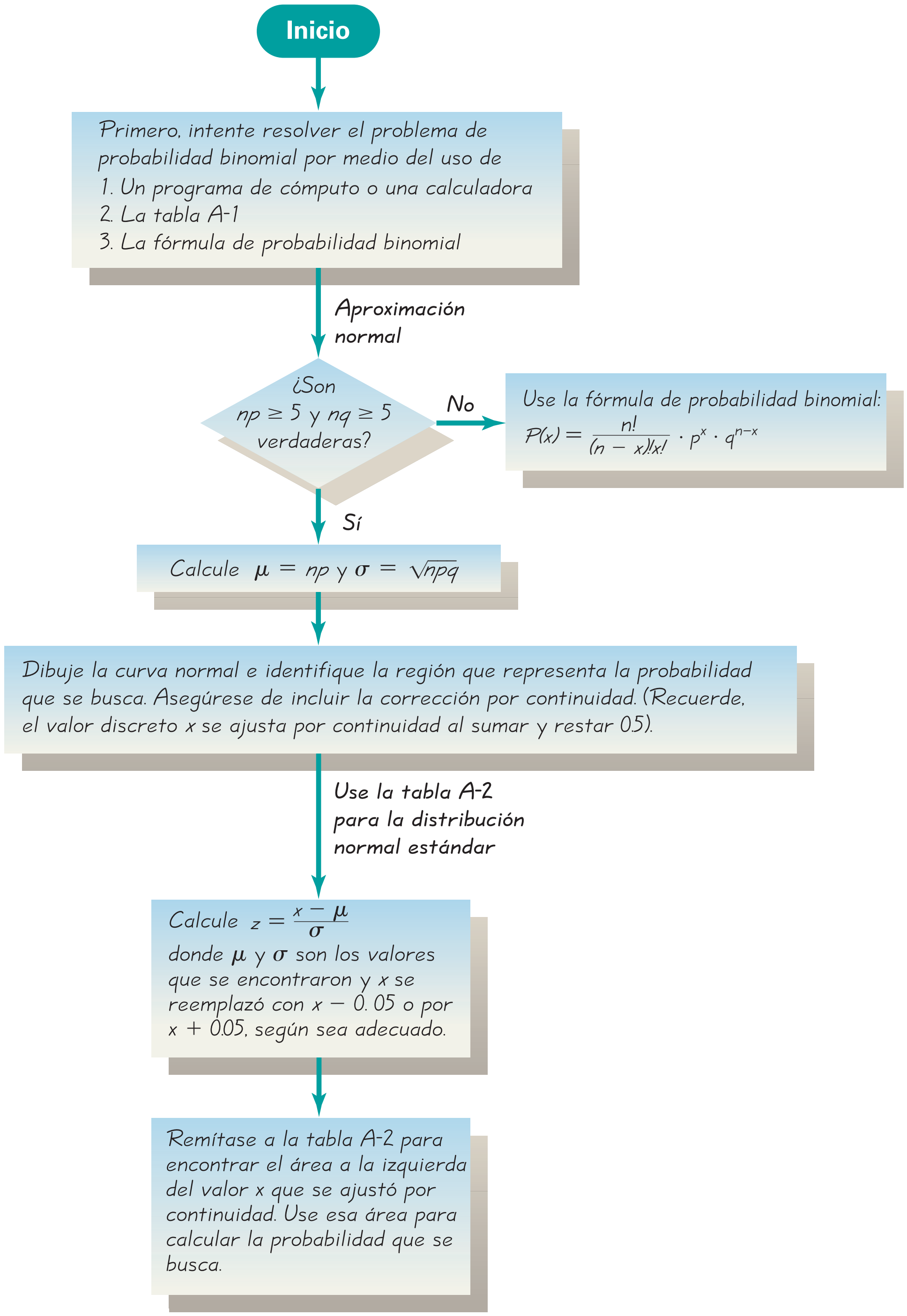


FIGURA 5-23 **Uso de la distribución normal como aproximación de la distribución binomial**

Para comprender mejor la manera en que la distribución normal, puede usarse para aproximar una distribución binomial, remítase a la figura 5-18 de la sección 5-4. Esta figura presenta un histograma de frecuencias relativas de los valores de 10,000 *proporciones* de muestra, donde cada una de las 10,000 muestras consta de 50 géneros que se seleccionaron al azar con reemplazo, de una población en la que la proporción de mujeres es de 0.13. Dichas proporciones muestrales pueden considerarse probabilidades binomiales, de manera que la figura 5-18 indica que, en condiciones adecuadas, las probabilidades binomiales tienen una distribución muestral aproximadamente normal. La justificación formal que nos permite emplear la distribución normal como aproximación de la distribución binomial resulta de matemáticas más avanzadas; la figura 5-18 es un argumento visual convincente que apoya esa aproximación.

Al resolver problemas de probabilidad binomial, primero intente obtener resultados más exactos por medio de un programa de cómputo, una calculadora, la tabla A-1 o la fórmula de probabilidad binomial. Si la probabilidad binomial no puede calcularse con estos procedimientos más exactos, intente la técnica del uso de la distribución normal como aproximación de la distribución binomial. Este método implica el siguiente procedimiento, que también se presenta en un diagrama de flujo en la figura 5-23.

Procedimiento para el uso de una distribución normal como aproximación de la distribución binomial

1. Establezca que la distribución normal es una aproximación adecuada de la distribución binomial, verificando que $np \geq 5$ y $nq \geq 5$. (Si ambas condiciones no se satisfacen, entonces debe utilizar un programa de cómputo, una calculadora, la tabla A-1 o la fórmula de probabilidad binomial).
2. Obtenga los valores de los parámetros μ y σ calculando $\mu = np$ y $\sigma = \sqrt{npq}$.
3. Identifique el valor discreto x (el número de éxitos). Reemplace el valor *discreto* x con el *intervalo* desde $x - 0.5$ hasta $x + 0.5$. (Para mayor explicación consulte la sección titulada “Correcciones por continuidad”, más adelante en esta sección). Dibuje una curva normal e introduzca los valores de μ, σ y $x - 0.5$ o $x + 0.5$, según sea apropiado.
4. Modifique x reemplazándola por $x - 0.5$ o $x + 5$, según sea apropiado.
5. Utilice $x - 0.5$ o $x + 0.5$ (según sea apropiado) en lugar de x, calcule el área correspondiente a la probabilidad deseada encontrando primero la puntuación z: $z = (x - \sigma\,)/\sigma$. Ahora use esa puntuación z para encontrar el área a la izquierda de $x - 0.5$ o $x + 0.5$, según sea apropiado. Ahora el área puede emplearse para identificar el área correspondiente a la probabilidad que se desea.

Ilustraremos este procedimiento de aproximación normal con el siguiente ejemplo.

EJEMPLO Cargas de aviones Cuando un avión se carga con pasajeros, equipaje, carga y combustible, el piloto debe verificar que el peso completo no rebase el límite máximo que se permite, por lo cual el peso tiene que distribuirse de forma conveniente para que el equilibrio del avión esté dentro de

continúa

Los niños y las niñas no son igualmente probables

En muchos cálculos de probabilidad, se obtienen buenos resultados suponiendo que los niños y las niñas tienen la misma posibilidad de nacer. En realidad, un niño tiene mayores posibilidades de nacer (probabilidad de 0.5117) que una niña (probabilidad de 0.4883). Tales resultados se basan en datos recientes del National Center for Health Statistics de Estados Unidos, institución que mostró que los 4,058,814 nacimientos que ocurrieron en un año incluyeron 2,076,969 niños y 1,981,845 niñas. Los investigadores están pendientes de dichas probabilidades por los cambios que podrían sugerir factores; entre otros, en el ambiente y la exposición a agentes químicos.

LA ESTADÍSTICA EN LAS NOTICIAS

Ganadores múltiples de la lotería

Evelyn Marie Adams ganó la lotería de Nueva Jersey en dos ocasiones en cuatro meses. Este feliz suceso fue reportado en los medios como una increíble coincidencia con tan sólo una posibilidad en 17 billones. Sin embargo, los matemáticos Persi Diaconis y Frederick Mosteller, de Harvard, señalan que hay una posibilidad en 17 billones de que una persona en particular, que posea un boleto para cada una de las dos loterías de Nueva Jersey, gane dos veces; pero existe aproximadamente una posibilidad en 30 de que alguien en Estados Unidos gane la lotería dos veces durante un periodo de cuatro meses. Diaconis y Mosteller analizaron las coincidencias y concluyeron que "con una muestra lo suficientemente grande, cualquier cosa sorprendente puede suceder". Según el *Detroit News*, Joe y Dolly Hornick ganaron la lotería de Pennsylvania cuatro veces en 12 años, con premios de $2.5 millones, $68,000; $206,217 y $71,037, respectivamente.

los límites aceptables de seguridad. Air America estableció un procedimiento según el cual debe reducirse la carga extra siempre que un avión con 200 pasajeros incluya al menos 120 hombres. Calcule la probabilidad de que, de 200 pasajeros seleccionados al azar, haya al menos 120 hombres. Suponga que la población de pasajeros potenciales consiste en un número igual de hombres y mujeres.

SOLUCIÓN Remítase a la figura 5-23 que presenta el procedimiento que se realizó para esta solución. El problema dado implica una distribución binomial con un número fijo de ensayos ($n = 200$), que se presume son independientes, con dos categorías de resultados (hombres, mujeres) para cada ensayo, y con la probabilidad de un hombre ($p = 0.5$) que se supone permanece constante de un ensayo a otro.

Supondremos que no disponemos de un programa de cómputo ni de una calculadora. La tabla A-1 no puede aplicarse, porque termina en $n = 15$. La fórmula de probabilidad binomial no es práctica, ya que tendríamos que utilizarla en 81 ocasiones (una para cada valor de x desde 120 hasta 200, inclusive), y nadie en su sano juicio desearía hacerlo.

Procedamos con el método de los cinco pasos para el uso de una distribución normal como aproximación de la distribución binomial.

Paso 1: Primero debemos verificar que es razonable aproximar la distribución binomial con la distribución normal, porque $np \geq 5$ y $nq \geq 5$. Con $n = 200$, $p = 0.5$ y $q = 1 - p = 0.5$, verificamos las condiciones requeridas como sigue:

$$np = 200 \cdot 0.5 = 100 \qquad \text{(Por lo tanto, } np \geq 5\text{).}$$
$$nq = 200 \cdot 0.5 = 100 \qquad \text{(Por lo tanto, } nq \geq 5\text{).}$$

Paso 2: Ahora procedamos a calcular los valores de μ y σ, necesarios para la distribución normal. Obtendremos lo siguiente:

$$\mu = np = 200 \cdot 0.5 = 100$$
$$\sigma = \sqrt{npq} = \sqrt{200 \cdot 0.5 \cdot 0.5} = 7.0710678$$

Paso 3: El valor discreto de 120 se representa con la franja que se limita con 119.5 y 120.5. (Véase la explicación sobre correcciones por continuidad, que sigue de este ejemplo).

Paso 4: Ya que buscamos la probabilidad de *al menos* 120 hombres, queremos el área que representa el número discreto de 120 (la región limitada por 119.5 y 120.5), así como también el área a la derecha, como se muestra en la figura 5-24.

Paso 5: Ahora es posible proceder a la búsqueda del área que se sombreó de la figura 5-24, utilizando los mismos métodos que se emplearon en la sección 5-3. Para usar la tabla A-2 de la distribución normal estándar, primero habrá que transformar 119.5 a una puntuación z; después, usar la tabla para encontrar el área a la izquierda de 119.5, que posteriormente se resta de 1. La puntuación z se obtiene como sigue:

$$z = \frac{x - \mu}{\sigma} = \frac{119.5 - 100}{7.0710678} = 2.76$$

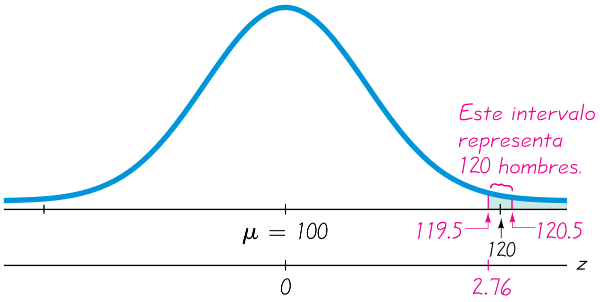


FIGURA 5-24 Búsqueda de la probabilidad de "al menos" 120 hombres entre 200 pasajeros

Al emplear la tabla A-2, encontramos que $z = 2.76$ corresponde a un área de 0.9971, de manera que la región que se sombreó es de $1 - 0.9971 = 0.0029$.

INTERPRETACIÓN Hay una probabilidad de 0.0029 de obtener al menos 120 hombres entre 200 pasajeros. Como esa probabilidad es muy baja, concluimos que, en muy pocas ocasiones, una lista de 200 pasajeros incluirá al menos 120 hombres, por lo que no es necesario preocuparse mucho por reducir la carga extra.

Correcciones por continuidad

El procedimiento que implica el uso de la distribución normal como aproximación de la distribución binomial incluye un paso en el que cambiamos un número discreto por un intervalo que está 0.5 por abajo y 0.5 por arriba del número discreto. Observe la solución anterior, donde cambiamos 120 por el intervalo entre 119.5 y 120.5. Este paso particular, que se denomina *corrección por continuidad*, suele ser difícil de comprender, por lo que ahora lo explicaremos con mayor detalle.

Definición

Cuando empleamos la distribución normal (que es una distribución de probabilidad *continua*) como un una aproximación de la distribución binomial (que es *discreta*), se realiza una **corrección por continuidad** a un número entero discreto x en la distribución binomial, representando el valor único x en el *intervalo* desde $x - 0.5$ hasta $x + 0.5$ (es decir, sumando y restando 0.5).

Las siguientes sugerencias prácticas deben ayudarlo a utilizar las correcciones por continuidad en forma apropiada.

Procedimiento para correcciones por continuidad

1. Cuando use la distribución normal como aproximación de la distribución binomial, *siempre* aplique la corrección por continuidad. (Esto se requiere, porque estamos utilizando la distribución normal *continua* para aproximar la distribución binomial *discreta*).
2. Para emplear la corrección por continuidad, primero identifique el número entero discreto x relevante al problema de probabilidad binomial. Por ejemplo,

si usted está intentando calcular la probabilidad de obtener al menos 120 hombres entre 200 personas que se seleccionaron aleatoriamente, el número entero discreto relevante sería $x = 120$. Primero enfoque su atención en el valor x e ignore temporalmente si busca al menos x, más que x, menos que x, u otro.

3. Dibuje una distribución normal centrada alrededor de μ; después, una *franja vertical* alrededor de x. Marque el lado izquierdo de la franja con el número igual a $x - 0.5$ y el lado derecho con el número igual a $x + 0.5$. Por ejemplo, para $x = 120$, dibuje una franja desde 119.5 hasta 120.5. *Considere el área completa de la franja para representar la probabilidad del número discreto x.*
4. Ahora determine si el valor de x debe incluirse en la probabilidad que busca. (Por ejemplo, "al menos x" incluye a x, pero "más que x" no la incluye). Después, determine si busca la probabilidad de al menos x, a lo sumo x, más que x, menos que x o exactamente x. Sombree el área a la derecha o a la izquierda de la franja, según sea apropiado; también el interior de la franja *si y sólo si* x se incluirá. Esta región total que se sombreó corresponde a la probabilidad buscada.

Para ver cómo resulta este procedimiento en las correcciones por continuidad, observe los casos comunes que se ilustran en la figura 5-25. Esos casos corresponden a los enunciados de la siguiente lista.

Enunciado	**Área**
Al menos 120 (incluye 120 y números mayores)	A la *derecha* de 119.5
Más de 120 (no incluye 120)	A la *derecha* de 120.5
A lo sumo 120 (incluye 120 y números menores)	A la *izquierda* de 120.5
Menos de 120 (no incluye 120)	A la *izquierda* de 119.5
Exactamente 120	Entre 119.5 y 120.5

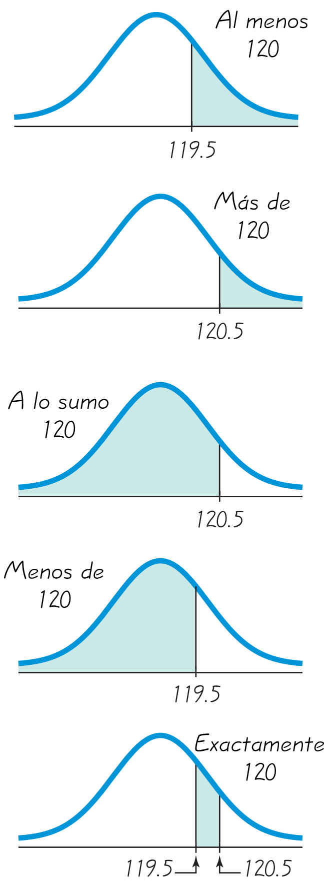


FIGURA 5-25 Uso de las correcciones por continuidad

EJEMPLO Audiencia televisiva El programa de televisión *60 minutos*, de la CBS, tuvo recientemente una audiencia de 20, lo que significa que, de los televisores en uso, el 20% estaba sintonizando *60 minutos* (de acuerdo con datos de Nielsen Media Research). Un anunciante desea verificar este valor del 20% de audiencia realizando su propia encuesta a 200 hogares que tengan su televisión encendida en el momento de la transmisión de *60 minutos*. Los resultados indican que, de los 200 televisores en uso, el 16% (o 32 televisores) están sintonizando *60 minutos*. Suponiendo que el valor de audiencia del 20% sea correcto, calcule la probabilidad de que en una encuesta de 200 hogares, *exactamente* 32 televisores estén sintonizando *60 minutos*. Puesto que el resultado muestral del 16% es menor que el valor de audiencia que se anunció del 20%, ¿hay evidencia fuerte para concluir que el valor de audiencia del 20% es incorrecto?

SOLUCIÓN Tenemos $n = 200$ ensayos independientes, $x = 32$ televisores sintonizando *60 minutos* y una proporción poblacional de $p = 0.20$. Para los propósitos de este ejemplo, suponemos que no se nos permite el acceso a un programa de cómputo ni a una calculadora TI-83 Plus. Tampoco es posible utilizar la tabla A-1, porque $n = 200$ excede el valor más alto de la tabla de

$n = 15$. Si utilizamos la fórmula de probabilidad binomial, deberíamos evaluar una expresión que incluya 200!, pero muchas calculadoras y programas de cómputo no manejan tantos datos. Por consiguiente, procedemos a emplear una distribución normal para aproximar la distribución binomial.

Paso 1: Primero verificamos si es posible la aproximación:

$$np = 200 \cdot 0.20 = 40 \qquad (\text{Por lo tanto, } np \geq 5.)$$
$$nq = 200 \cdot 0.80 = 160 \qquad (\text{Por lo tanto, } nq \geq 5.)$$

Paso 2: Ahora procedemos a calcular los valores de μ y σ, necesarios para la distribución normal. Obtenemos lo siguiente:

$$\mu = np = 200 \cdot 0.20 = 40$$
$$\sigma = \sqrt{npq} = \sqrt{200 \cdot 0.20 \cdot 0.80} = 5.6568542$$

Paso 3: Dibujamos la curva normal de la figura 5-26. La región sombreada de la figura representa la probabilidad que buscamos. La aplicación de la corrección por continuidad da como resultado la representación de 32, ubicada entre 31.5 y 32.5.

Paso 4: He aquí el método que se empleó para calcular la región sombreada de la figura 5-26: primero calcule el área total a la izquierda de 32.5; después, obtenga el área total a la izquierda de 31.5; luego, calcule la *diferencia* entre ambas áreas. Iniciando con el área total a la izquierda de 32.5, debemos obtener la puntuación z que corresponde a 32.5. Si nos remitimos a la tabla A-2, obtendremos

$$z = \frac{32.5 - 40}{5.6568542} = -1.33$$

Usamos la tabla A-2 para encontrar que $z = -1.33$ corresponde a una probabilidad de 0.0918, que es el área total a la izquierda de 32.5. Ahora, procedemos a obtener el área a la izquierda de 31.5, calculando primero la puntuación z correspondiente a 31.5:

$$z = \frac{31.5 - 40}{5.6568542} = -1.50$$

En la tabla A-2 encontramos que $z = -1.50$ corresponde a una probabilidad de 0.0668, que es el área total a la izquierda de 31.5. El área sombreada es $0.0918 - 0.0668 = 0.0250$.

continúa

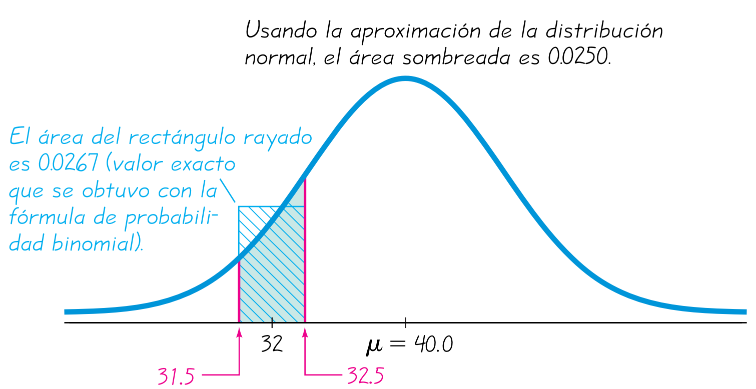


FIGURA 5-26 Uso de la corrección por continuidad de la audiencia televisiva

INTERPRETACIÓN La probabilidad de que exactamente 32 televisores sintonicen *60 minutos* (de un total de 200) es de aproximadamente 0.0250. El planteamiento del problema también nos pide determinar si el resultado muestral del 16% constituye una evidencia suficiente para concluir que el valor de audiencia del 20% es incorrecto. Sin embargo, en lugar de considerar la probabilidad de *exactamente* 32 televisores que sintonizan *60 minutos*, debemos considerar la probabilidad de *32 o menos*. [En la sección 4-2, señalamos que x éxitos en n ensayos es un número de éxitos *infrecuentemente bajo*, si $P(x$ o menor) es muy pequeña, como 0.05 o menos]. En la solución anterior vemos que la probabilidad de 32 o menos éxitos es P(menos que 32.5), que es 0.0918. Puesto que 0.0918 no es muy pequeño, *no* tenemos suficiente evidencia para concluir que el valor de audiencia del 20% sea incorrecto.

Si resolvemos el ejemplo anterior por medio de STATDISK, Minitab o una calculadora, obtendremos un resultado de 0.0267, pero el método de aproximación normal arrojó un valor de 0.0250. La discrepancia de 0.0017 sucede porque el uso de la distribución normal da como resultado un valor que se *aproxima* al que corresponde al área de la región sombreada en la figura 5-26, mientras que el área correcta exacta es un rectángulo que se centra por arriba de 32. (La figura 5-26 ilustra tal discrepancia). El área del rectángulo es 0.0267, pero el área aproximada de la región sombreada es 0.0250.

Interpretación de los resultados

En realidad, cuando utilizamos una distribución normal como aproximación de la distribución binomial, nuestra meta no es sencillamente calcular un número de probabilidad. Con frecuencia necesitamos hacer algún *juicio* con base en el valor de probabilidad, como en la conclusión final del ejemplo anterior. Debemos comprender que las bajas probabilidades corresponden a sucesos con pocas posibilidades, mientras que las altas probabilidades corresponden a sucesos posibles. El valor de probabilidad de 0.05 suele utilizarse como punto de corte para distinguir entre sucesos posibles y sucesos imposibles. El siguiente criterio (de la sección 4-2) describe la aplicación de las probabilidades para distinguir resultados que pueden ocurrir fácilmente por el azar, de aquellos que son en extremo poco comunes.

Uso de las probabilidades para determinar cuando los resultados son poco comunes

- **Extremadamente alto:** x éxitos en n ensayos es un número *extremadamente alto* de éxitos si $P(x$ o más) es muy pequeña (como 0.05 o menos).
- **Extremadamente bajo:** x éxitos en n ensayos es un número *extremadamente bajo* de éxitos si $P(x$ o menos) es muy pequeña (como 0.05 o menos).

5-6 Destrezas y conceptos básicos

Aplicación de la corrección por continuidad. *En los ejercicios 1 a 8, los valores dados son discretos. Utilice la corrección por continuidad y describa la región de la distribución normal que corresponde a la probabilidad que se indica. Por ejemplo, la probabilidad de "más que 20 artículos defectuosos" corresponde al área de la curva normal descrita en esta respuesta: "el área a la derecha de 20.5".*

1. Probabilidad de que más de 15 personas en prisión quiten las etiquetas de advertencia a las almohadas.
2. Probabilidad de que al menos 24 estudiantes comprendan la corrección por continuidad.
3. Probabilidad de que haya menos de 100 pasajeros en su siguiente vuelo comercial.
4. Probabilidad de que el número de distribuidores automáticos en Estados Unidos sea exactamente 27.
5. Probabilidad de no más de cuatro estudiantes ausentes en una clase de estadística.
6. Probabilidad de que el número de CD defectuosos de Wayne Newton sea de 15 a 20, inclusive.
7. Probabilidad de que el número de senadores estadounidenses ausentes sea de ocho a 10, inclusive.
8. Probabilidad de exactamente tres respuestas "sí" en peticiones de citas.

Uso de la aproximación normal. *En los ejercicios 9 a 12, haga lo siguiente: a) Calcule la probabilidad binomial que se indica por medio de la tabla A-1 del Apéndice A. b) Si* $np \geq 5$ *y* $nq \geq 5$*, también estime la probabilidad que se indica con el uso de la distribución normal como aproximación de la distribución normal; si* $np < 5$ *o* $nq < 5$*, entonces establezca que la aproximación normal no es adecuada.*

9. Con $n = 14$ y $p = 0.5$, calcule $P(9)$.
10. Con $n = 12$ y $p = 0.8$, calcule $P(7)$.
11. Con $n = 15$ y $p = 0.9$, calcule P(al menos 14).
12. Con $n = 13$ y $p = 0.4$, calcule P(menor que 3).
13. Probabilidad de más de 55 niñas Estime la probabilidad de que resulten más de 55 niñas en 100 nacimientos. Suponga que los niños y las niñas son igualmente probables. ¿Es poco común que resulten más de 55 niñas en 100 nacimientos?
14. Probabilidad de al menos 65 niñas Estime la probabilidad de que resulten al menos 65 niñas en 100 nacimientos. Suponga que los niños y las niñas son igualmente probables. ¿Es poco común que resulten al menos 65 niñas en 100 nacimientos?
15. Probabilidad de al menos aprobar Estime la probabilidad de aprobar un examen de verdadero/falso de 100 preguntas, si el 60% (o 60 respuestas correctas) es la calificación mínima de aprobación y si todas las respuestas son conjeturas. ¿Es la probabilidad lo suficientemente alta como para arriesgarse a aprobar adivinando en lugar de estudiar?
16. Examen de opción múltiple Un examen de opción múltiple consta de 25 preguntas con las respuestas posibles a, b, c, d y e. Estime la probabilidad de que, al adivinar, el número de respuestas correctas sea de tres a 10, inclusive.
17. Experimento de hibridación de Mendel Cuando Mendel realizó sus famosos experimentos de hibridación, utilizó chícharos con vainas verdes y vainas amarillas. Uno de los experimentos implicó una cruza de chícharos, de manera que se esperaba que el 25% (o 145) de los 580 chícharos vástagos tuvieran vainas amarillas. En lugar de obtener 145 chícharos con vainas verdes, obtuvo 152. Suponiendo que el porcentaje del 25% de Mendel es correcto, estime la probabilidad de obtener al menos 152 chícharos con vainas amarillas, entre los 580 chícharos vástagos. ¿Existirá una fuerte evidencia que sugiera que la probabilidad del 25% de Mendel es incorrecta?
18. Fármaco que reduce el colesterol La probabilidad de que una persona que no recibe ningún tratamiento tenga síntomas de gripe es de 0.019. En un ensayo clínico de Lipitor, un fármaco común que se utilizó para disminuir el colesterol, 863 pacientes recibieron un tratamiento con tabletas de Atorvastatin de 10 mg, y 19 de estos pacientes experimentaron síntomas de gripe (según datos de Pfizer, Inc.). Suponiendo que estas tabletas no influyen en los síntomas de la gripe, estime la probabilidad de que al menos 19 de

las 863 personas experimenten síntomas de gripe. ¿Sugieren estos resultados acerca de los síntomas de gripe que hay una reacción adversa al fármaco?

19. Probabilidad de al menos 50 hombres daltónicos El 9% de los hombres y el 0.25% de las mujeres no pueden distinguir entre los colores rojo y verde. Este tipo de daltonismo causa problemas con las señales de tránsito. Los investigadores necesitan al menos 50 hombres con este tipo de ceguera al color, de manera que seleccionan aleatoriamente a 600 hombres para un estudio de percepción de las señales de tránsito. Estime la probabilidad de que al menos 50 de los hombres no distingan entre el rojo y el verde. ¿Es el resultado lo suficientemente alto como para qué los investigadores puedan confiarse de obtener al menos 50 hombres con daltonismo?

20. Teléfonos celulares y cáncer cerebral En un estudio de 420,000 usuarios de teléfono celular en Dinamarca, se encontró que 135 desarrollaron cáncer cerebral o del sistema nervioso. Suponiendo que los teléfonos celulares no tienen efecto alguno, hay una probabilidad de 0.000340 de que una persona desarrolle cáncer cerebral o del sistema nervioso. Por lo tanto, esperaríamos aproximadamente 143 casos de este tipo de cáncer en un grupo de 420,000 personas seleccionadas al azar. Estime la probabilidad de 135 o menos casos de este cáncer en un grupo de 420,000 personas. ¿Qué sugieren estos resultados acerca de los reportes de los medios de comunicación que afirman que los teléfonos celulares causan cáncer cerebral o del sistema nervioso?

21. Vuelos sobresaturados Air America está considerando la nueva política de registrar 400 personas en un avión que tiene sólo 350 asientos. (Estudios anteriores han revelado que sólo el 85% de los pasajeros registrados llegan al vuelo). Estime la probabilidad de que, si Air America registra a 400 personas, no haya suficientes asientos disponibles. ¿Es esta probabilidad lo suficientemente baja para ser funcional, o deberá modificarse la política?

22. Vuelos a tiempo Recientemente, el 72.3% de los vuelos de American Airlines llegaron a tiempo (según datos del Departamento del Transporte de Estados Unidos). Al verificar 40 vuelos de American Airlines, seleccionados al azar, 19 llegaron a tiempo. Estime la probabilidad de que 19 vuelos o menos, entre 40, lleguen a tiempo suponiendo que el porcentaje del 72.3% sea correcto. ¿Será poco común que 19 vuelos o menos, entre 40 vuelos de American Airlines seleccionados aleatoriamente lleguen a tiempo?

23. Identificación de discriminación por género Después de que la rechazaron para un empleo, Kim Kelly se entera de que la Bellevue Advertising Company contrató únicamente a 21 mujeres entre sus 62 empleados nuevos. También de que el grupo de solicitantes es muy grande, con igual número de hombres y mujeres calificados. Ayúdela a hacer una acusación por discriminación, estimando la probabilidad de obtener 21 mujeres o menos cuando se contrata a 62 personas, suponiendo que no hay discriminación por género. ¿En realidad apoya la probabilidad resultante una acusación como ésta?

24. Dulces M&M: ¿el 10% son azules? Según un representante de asuntos de consumo de Mars (la compañía de dulces), el 10% de todos los dulces sencillos M&M son azules. El conjunto de datos 19 del Apéndice B indica que de 100 M&M elegidos, cinco son azules. Estime la probabilidad de seleccionar al azar 100 dulces M&M y obtener cinco o menos que sean azules. Suponga que el porcentaje de azules del 10%, establecido por la compañía, es correcto. Con base en el resultado, ¿será poco común obtener cinco o menos M&M azules cuando se seleccionan 100 al azar?

25. Grupo sanguíneo El 45% de nosotros tiene sangre del grupo O, según datos que proporcionó el Great New York Program. El Providence Memorial Hospital está realizando una campaña de donación de sangre, ya que su abastecimiento de sangre del grupo O es bajo y necesita 177 donadores de este tipo de sangre. Si 400 voluntarios donan sangre, estime la probabilidad de que el número de personas con sangre del grupo O sea al menos de 177. ¿Es probable que el grupo de 400 voluntarios sea suficiente?

26. Muestreo de aceptación En la sección 3-4 establecimos que algunas compañías verifican la calidad a través del método del muestreo de aceptación, por medio del cual se rechaza el lote completo de artículos si una muestra aleatoria de un tamaño particular incluye más de un número específico de defectos. La Dayton Machine Company compra tornillos de máquina en lotes de 5000 y rechaza un lote si, cuando se saca una muestra de 50, al menos dos son defectuosos. Estime la probabilidad de rechazar un lote si el abastecedor está fabricando los tornillos con una tasa de defectos del 10%. ¿Es posible que el plan de verificación identifique la tasa inaceptable de defectos?

27. Choques de automóviles Entre los conductores de 20 a 24 años de edad hay una tasa del 34% de accidentes automovilísticos en un año (según datos del National Safety Council de Estados Unidos). Un investigador de seguros encuentra que en un grupo de 500 conductores con edades que fluctúan entre 20 y 24 años, que se seleccionó aleatoriamente, y que viven en la ciudad de Nueva York, el 40% tuvo accidentes el año anterior. Si el porcentaje del 34% es correcto, estime la probabilidad de que en un grupo de 500 conductores seleccionados al azar, al menos el 40% tuvieran accidentes el año anterior. Con base en el resultado, ¿existe fuerte evidencia que apoye la aseveración de que la tasa de accidentes en la ciudad de Nueva York es mayor al 34%?

28. Encuesta sobre clonación Una reciente encuesta de Gallup incluyó 1012 adultos que se seleccionaron al azar, a quienes se les preguntó si "la clonación humana debe o no permitirse". Los resultados mostraron que el 89% de los encuestados indicaron que no debe permitirse. Un reportero de noticias desea determinar si estos resultados de encuesta constituyen una fuerte evidencia de que la mayoría (más del 50%) de las personas se oponen a dicha clonación. Suponiendo que el 50% de todas las personas se oponga, estime la probabilidad de obtener al menos 89% de oposición en una encuesta de 1012 personas seleccionadas al azar. Con base en el resultado, ¿hay fuerte evidencia que apoye la afirmación de que la mayoría se opone a la clonación de humanos?

5-6 Más allá de lo básico

29. Ganar en la ruleta Marc Taylor planea hacer 200 apuestas, de $1 cada una, al número 7 en la ruleta. Un triunfo paga con posibilidades de 35:1 y, en cualquier giro, existe una probabilidad de $1/38$ de que el 7 sea el número ganador. De las 200 apuestas, ¿cuál es el número mínimo de triunfos necesarios para que Marc obtenga una ganancia? Estime la probabilidad de que Marc obtenga una ganancia.

30. Reemplazo de televisores Los tiempos de reemplazo de televisores se distribuyen normalmente, con una media de 8.2 años y una desviación estándar de 1.1 años (de acuerdo con datos de "Getting Fixed", *Consumer Reports*). Estime la probabilidad de que, para 250 televisores seleccionados al azar, al menos 15 de ellos tengan tiempos de reemplazo mayores de 10.0 años.

81. 'Joltin' Joe Suponga que un jugador de beisbol pega de "hit" .350, de manera que su probabilidad de un "hit" es de 0.350. (Ignore las complicaciones causadas por las bases por bolas). También suponga que sus intentos de "hit" son independientes unos de otros.

a. Calcule la probabilidad de al menos un "hit" en cuatro intentos, en un juego.

b. Suponiendo que este bateador pasa a batear cuatro veces cada juego, estime la probabilidad de obtener un total de al menos 56 "hits" en 56 juegos.

c. Suponiendo que este bateador pasa a batear cuatro veces cada juego, estime la probabilidad de al menos un "hit" en cada uno de 56 juegos consecutivos (que es el récord de Joe DiMaggio en 1941).

d. ¿Cuál es el promedio mínimo de bateo que se requeriría para que la probabilidad del inciso *c* sea mayor que 0.1?

32. Vuelos sobresaturados Vertigo Airlines trabaja únicamente con reservaciones anticipadas y registra una tasa del 7% de personas que no se presentan. ¿Cuántas reservaciones podrían aceptarse para un avión con una capacidad de 250, si hay al menos una probabilidad de 0.95 de que a todos los individuos que reservaron y se presenten se les acomode?

5-7 Determinación de la normalidad

Los siguientes capítulos incluyen algunos métodos estadísticos muy importantes que requieren que los datos muestrales se seleccionen aleatoriamente a partir de una población con una distribución *normal*. Por consiguiente, es necesario determinar si los datos muestrales parecen provenir de una población distribuida normalmente. En esta sección introducimos la *gráfica cuantilar normal* como una herramienta que nos ayuda a determinar si aparentemente se satisfacen los requisitos de una distribución normal.

Definición

Una **gráfica cuantilar normal** es una gráfica de puntos (x, y) donde cada valor x proviene del conjunto original de datos muestrales, así como cada valor y es una puntuación z correspondiente a un valor cuantilar de la distribución normal estándar. (Véase el paso 3 en el siguiente procedimiento para conocer detalles sobre el cálculo de estas puntuaciones z).

Procedimiento para determinar si los datos se distribuyen normalmente

1. *Histograma:* Construya un histograma. Rechace la normalidad si el histograma difiere mucho de la forma de campana.
2. *Datos distantes:* Identifique datos distantes. Rechace la normalidad si hay más de un dato distante presente. (La presencia de un solo dato distante podría ser un error o el resultado de la variación por el azar, pero tenga cuidado porque incluso un solo dato distante llega a producir un efecto importante en los resultados).
3. *Gráfica cuantilar normal:* Si el histograma es básicamente simétrico y existe a lo sumo un dato distante, construya una *gráfica cuantilar normal*. Los siguientes pasos describen la construcción de una gráfica cuantilar normal, pero el procedimiento es tan confuso que solemos utilizar un programa de cómputo o una calculadora para generar la gráfica. Al final de esta sección se incluyen instrucciones para el uso de STATDISK, Minitab, Excel y la calculadora TI-83 Plus, para obtener gráficas cuantilares normales.
   - **a.** Primero ordene los datos del más bajo al más alto.
   - **b.** Con una muestra de tamaño n, cada valor representa una proporción de $1/n$ de la muestra. Utilizando el tamaño muestral n que se conoce, identifique las áreas de $1/2n$, $3/2n$, $5/2n$, $7/2n$, etcétera. Éstas son las áreas acumulativas a la izquierda de los valores muestrales correspondientes.
   - **c.** Utilice la distribución normal estándar (tabla A-2) para calcular las puntuaciones z correspondientes a las áreas izquierdas acumulativas que se obtuvieron en el paso *b*.
   - **d.** Una los valores originales de los datos ordenados con sus puntuaciones z correspondientes, que se calcularon en el paso *c*, después grafique los puntos (x, y), donde cada x es un valor muestral original, en tanto y es la puntuación z correspondiente.
   - **e.** Examine la gráfica cuantilar normal con los siguientes criterios: si los puntos no se acercan a una línea recta o si exhiben algún patrón sistemático

diferente al de una línea recta, entonces parece que los datos provienen de una población que *no* tiene una distribución normal. Si el patrón de puntos se acerca razonablemente a una línea recta, entonces los datos pueden provenir de una población con distribución normal.

Los pasos 1 y 2 son directos, pero ilustramos la construcción de una gráfica cuantilar normal (paso 3) en el siguiente ejemplo.

EJEMPLO Edades de presidentes Los ejercicios 8 y 13 de la sección 5-4 incluyen las edades de cinco presidentes de Estados Unidos con profesiones militares en el momento de tomar posesión: 62, 46, 68, 64, 57. Construya una gráfica cuantilar normal para las edades y determine si parecen provenir de una población que se distribuye normalmente.

SOLUCIÓN Los siguientes pasos corresponden a los listados en el procedimiento anterior para la construcción de una gráfica cuantilar normal.

1. Primero hay que ordenar los datos: 46, 57, 62, 64, 68.
2. Con una muestra de tamaño $n = 5$, cada valor representa una proporción de $1/5$ de la muestra, por lo que procedemos e identificar las áreas acumulativas a la izquierda de los valores muestrales correspondientes. Estas áreas izquierdas acumulativas, que se expresan en general como $1/2n$, $3/2n$, $5/2n$, $7/2n$, etcétera, se convierten en áreas específicas para el presente ejemplo, con $n = 5$: $1/10$, $3/10$, $5/10$, $7/10$ y $9/10$. Tales áreas izquierdas acumulativas, que se expresan en forma decimal, son 0.1, 0.3, 0.5, 0.7 y 0.9.
3. Ahora buscamos en la tabla A-2 las áreas izquierdas acumulativas de 0.1000, 0.3000, 0.5000, 0.7000 y 0.9000. Encontramos estas puntuaciones z correspondientes: -1.28, -0.52, 0, 0.52 y 1.28.
4. Ahora unimos las edades ordenadas con sus puntuaciones z correspondientes; obtenemos las siguientes coordenadas (x, y), que están graficadas en la figura 5.27: $(46, -1.28)$, $(57, -0.52)$, $(62, 0)$, $(64, 0.52)$ y $(68, 1.28)$.

INTERPRETACIÓN Examinamos la gráfica cuantilar normal de la figura 5-27. Como los puntos parecen estar razonablemente cerca de una línea recta, concluimos que las edades dadas parecen provenir de una población que se distribuye normalmente.

Puesto que la construcción de una gráfica cuantilar normal requiere que ordenemos los datos muestrales y que luego hagamos un proceso complicado para calcular las puntuaciones z correspondientes, entonces la construcción manual de la gráfica es

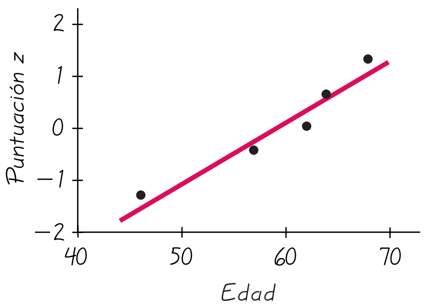


FIGURA 5-27 Gráfica normal cuantilar de las edades de presidentes

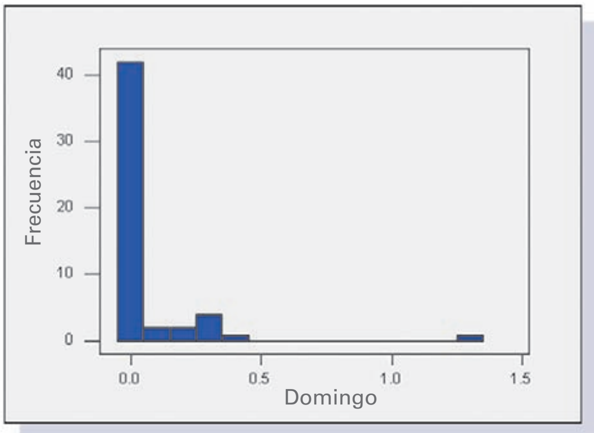
difícil con conjuntos grandes de datos. El siguiente ejemplo ilustra el uso de programas de cómputo.

EJEMPLO Lluvia en Boston En el conjunto de datos 11 del Apéndice B, utilice las 52 cantidades de lluvia de los domingos en Boston y haga una prueba de normalidad.

SOLUCIÓN

Paso 1: Construya un histograma. La siguiente pantalla de Minitab incluye el histograma de las 52 cantidades de lluvia, el cual presenta un sesgo extremo, lo que sugiere que dichas cantidades no se distribuyen de manera normal.

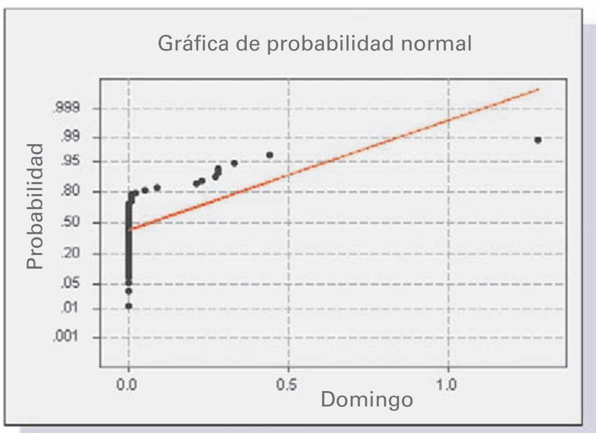
Minitab **Histograma de las cantidades de lluvia de los domingos en Boston**



Paso 2: Identifique datos distantes. Si examinamos la lista de 52 cantidades de lluvia, encontramos que 1.28 pulgadas parece ser el único dato distante. Debido a que sólo hay un dato distante, no sacamos conclusiones sobre la normalidad de los datos, con base en los datos distantes.

Paso 3: Construya una gráfica cuantilar normal. La siguiente pantalla de Minitab incluye una gráfica *de probabilidad* normal. (Puesto que muchos

Minitab **Gráfica de probabilidad normal de las cantidades de lluvia de los domingos en Boston**



valores de datos son iguales, la gráfica de probabilidad normal original incluye únicamente 12 puntos distintos en lugar de 52, de manera que la gráfica se modificó para mostrar los 52 puntos). La gráfica de probabilidad normal es igual a la gráfica cuantilar normal, excepto por la escala del eje vertical. Una gráfica de probabilidad normal se interpreta con los mismos criterios que una gráfica cuantilar normal. El examen de una gráfica de probabilidad normal revela un patrón muy diferente al patrón de una línea recta, lo que sugiere que los datos no provienen de una población distribuida normalmente.

INTERPRETACIÓN Puesto que el histograma no parece tener forma de campana, y porque la gráfica de probabilidad normal no produce un patrón de puntos que se aproxime razonablemente a una línea recta, concluimos que las cantidades de lluvia en Boston los domingos no se distribuyen de manera normal. Algunos de los procedimientos estadísticos en los capítulos posteriores requieren que los datos muestrales se distribuyan de manera normal, pero ese requisito no se satisface para las cantidades de lluvia de Boston los domingos, por lo que tales procedimientos no pueden aplicarse.

A continuación presentamos unos comentarios finales acerca de los procedimientos que se emplean para determinar si los datos provienen de una población distribuida de manera normal:

- Si el requisito de una distribución normal no es muy estricto, el examen de un histograma y de los datos distantes podría ser todo lo que necesite para determinar la normalidad.
- Las gráficas cuantilares normales en ocasiones resultan difíciles de construir, pero pueden generarse con una calculadora TI-83 Plus o con un programa de cómputo como STATDISK, Minitab y Excel.
- Además de los procedimientos estudiados en esta sección, hay otros procedimientos más avanzados, como la chi cuadrada, la prueba de bondad de ajuste, la prueba de Kolmogorov-Smirnov y la prueba de Lilliefors. (Véase "Beyond Basic Statistics with the Graphing Calculator, Part I: Assessing Goodness-of-fit", de Calzada y Scariano, *Mathematics and Computer Education*).

Utilizando la tecnología

STATDISK STATDISK puede utilizarse para generar una gráfica cuantilar normal. Primero seleccione **Data** de la parte superior de la barra del menú principal, luego seleccione **Normal Quantile Plot.** Proceda a introducir los datos y haga clic en **Evaluate.**

Minitab Minitab puede emplearse para generar una *gráfica de probabilidad normal*, que se interpreta de la misma forma que la gráfica cuantilar normal. Es decir, los datos que se distribuyen de manera normal deben aproximarse a una línea recta. Primero introduzca los valores en la columna C1, después seleccione **Stat, Basic Statistics** y **Normality Test.** Introduzca **C1** para la variable, después haga clic en **OK**.

Excel El complemento Data Desk XL puede utilizarse para generar una *gráfica de probabilidad normal*, que se interpreta de la misma manera que una gráfica cuantilar normal. Primero

continúa

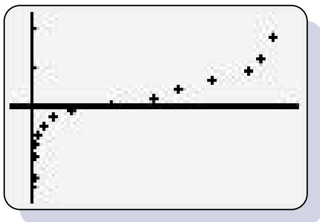
introduzca los valores muestrales en la columna A, después haga clic en **DDXL.** (Si DDXL no aparece en la barra del menú, instale el complemento Data Desk XL). Seleccione **Charts and Plots**, después seleccione la función de **Normal Probability Plot.** Haga clic en el icono del lápiz para “Quantitative Variable”, luego introduzca rangos de valores, tales como A1:A36. Presione **OK**.

TI-83 Plus La calculadora TI-83 Plus permite generar una gráfica cuantilar normal de la siguiente manera: primero introduzca los datos muestrales en la lista L1, presione **2nd** y la tecla **Y =** (para **STAT PLOT**), y después, **ENTER**. Seleccione **ON**, seleccione el elemento “type”, que es el último del segundo renglón de opciones, luego **L1** para la lista de datos. Luego de hacer todas las selecciones, presione **ZOOM** y luego **9**.

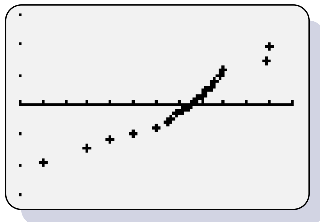
5-7 Destrezas y conceptos básicos

Interpretación de gráficas cuantilares normales. *En los ejercicios 1 a 4, examine la gráfica cuantilar normal y determine si describe datos que tienen una distribución normal.*

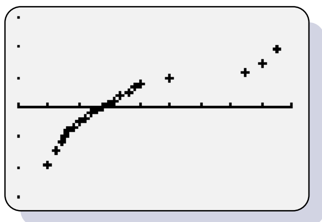
1.



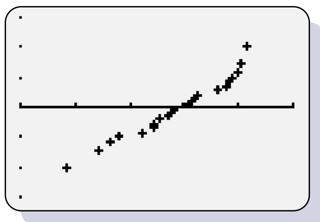
2.



3.



4.



Determinación de normalidad. *En los ejercicios 5 a 8, remítase al conjunto de datos que se indican y determine si se satisface el requisito de una distribución normal. Suponga que este requisito es flexible, en el sentido de que la distribución poblacional no necesita ser exactamente normal, sino que debe tratarse de una distribución que sea básicamente simétrica y con una moda única.*

5. Lluvia en Boston Las cantidades de lluvia que caen en Boston los miércoles, como se lista en el conjunto de datos 11 del Apéndice B.

6. Circunferencia de cabezas Las circunferencias de las cabezas de hombres, como se lista en el conjunto de datos 3 del Apéndice B.

7. Pesos de M&M Los pesos de los dulces M&M color café, como se lista en el conjunto de datos 19 del Apéndice B.

8. Conductividad del agua Los niveles de conductividad de los Everglades de Florida, como se lista en el conjunto de datos 12 del Apéndice B.

Generación de gráficas cuantilares normales. *En los ejercicios 9 a 12, utilice los datos del ejercicio que se indica en esta sección. Emplee una calculadora TI-83 Plus o un programa de cómputo (como STATDISK, Minitab o Excel), capaces de generar gráficas cuantilares normales o gráficas de probabilidad normal. Genere la gráfica y después determine si los datos provienen de una población distribuida normalmente.*

9. Ejercicio 5

10. Ejercicio 6

11. Ejercicio 7

12. Ejercicio 8

13. Comparación de conjuntos de datos Con las estaturas y los niveles de colesterol de mujeres, que se listan en el conjunto de datos 1 del Apéndice B, analice cada uno de los dos conjuntos de datos y determine si cada uno de ellos parece provenir de una población distribuida de manera normal. Compare los resultados y dé una posible explicación para cualquier diferencia notoria entre las dos distribuciones.

14. Comparación de conjuntos de datos Con los niveles de presión sanguínea histórica y las anchuras del codo de mujeres, que se listan en el conjunto de datos 1 del Apéndice B, analice cada uno de los dos conjuntos de datos y determine si cada uno de ellos parece provenir de una población distribuida de manera normal. Compare los resultados y dé una posible explicación para cualquier diferencia notoria entre las dos distribuciones.

Construcción de gráficas cuantilares normales. *En los ejercicios 15 y 16, utilice los valores dados e identifique las puntuaciones* z *correspondientes que se emplean para una gráfica cuantilar normal, después construya la gráfica cuantilar normal y determine si los datos parecen provenir de una población con una distribución normal.*

15. Estaturas de los Lakers de L.A. Utilice esta muestra de estaturas (en pulgadas) de los jugadores de la alineación estelar del equipo profesional de basquetbol de los Lakers de Los Ángeles: 85, 79, 82, 73, 78.

16. Monitoreo del plomo en el aire En los días siguientes a la destrucción que causaron los ataques terroristas del 11 de septiembre de 2001 se registraron las cantidades del plomo en el aire (en microgramos por metro cúbico), en el edificio 5 del World Trade Center, y se obtuvieron los siguientes valores: 5.40, 1.10, 0.42, 0.73, 0.48, 1.10.

5-7 Más allá de lo básico

17. Uso de puntuaciones estándar Al construir una gráfica cuantilar normal, suponga que en lugar de calcular las puntuaciones z por medio del procedimiento descrito en esta sección, cada valor en una muestra se transforma a su puntuación estándar correspondiente a través de $z = (x - \bar{x})/s$. Si los puntos (x, y) se marcan en una gráfica, ¿es posible usar esta gráfica para determinar si la muestra proviene de una población distribuida normalmente? Explique.

18. Distribución log normal Se considera que la variable aleatoria x tiene una *distribución log normal,* si los valores de ln x se distribuyen normalmente. Pruebe la normalidad de las siguientes duraciones de llamadas telefónicas (en segundos), después pruebe la normalidad de los logaritmos naturales de las duraciones. ¿Qué concluye?

31.5	75.9	31.8	87.4	54.1	72.2	138.1	47.9	210.6	127.7
160.8	51.9	57.4	130.3	21.3	403.4	75.9	93.7	454.9	55.1

Repaso

En el capítulo 4 estudiamos el concepto de distribuciones de probabilidad, pero sólo incluimos las distribuciones *discretas*. En este capítulo estudiamos las distribuciones de probabilidad *continua*, enfocándonos en su categoría más importante: las distribuciones normales. Las distribuciones normales se utilizarán continuamente en los siguientes capítulos.

Cuando se grafican, las distribuciones normales se aproximan a una forma de campana. El área total bajo la curva de densidad de una distribución normal es 1, de manera que hay una correspondencia conveniente entre áreas y probabilidades. Las áreas específicas pueden encontrarse por medio de la tabla A-2, de una calculadora TI-83 Plus o de un programa de cómputo. (No utilizamos la fórmula 5-1, que es la ecuación utilizada para definir la distribución normal).

En este capítulo presentamos métodos importantes para trabajar con las distribuciones normales, incluyendo las que emplean la puntuación estándar $z = (x - \mu)/\sigma$ para resolver problemas como éstos:

- Puesto que las puntuaciones de CI se distribuyen normalmente, con $\mu = 100$ y $\sigma = 15$, calcule la probabilidad de seleccionar aleatoriamente a un individuo con un CI por arriba de 90.
- Puesto que las puntuaciones de CI se distribuyen normalmente, con $\mu = 100$ y $\sigma = 15$, calcule la puntuación de CI que separa al 85% inferior del 15% superior.

En la sección 5-4 presentamos el concepto de distribución muestral. La distribución muestral de la media es la distribución de probabilidad de medias de muestra, donde todas las muestras tienen el mismo tamaño de muestra *n*. La distribución muestral de la proporción es la distribución de probabilidad de proporciones muestrales, donde todas las muestras tienen el mismo tamaño de muestra *n*. En general, la distribución muestral de cualquier estadístico es la distribución de probabilidad de dicho estadístico.

En la sección 5-5 presentamos los siguientes puntos importantes, que se asocian con el teorema del límite central:

1. La distribución de medias de muestra se aproxima a la distribución normal, conforme el tamaño de muestra *n* se incrementa.
2. La media de las medias de muestra es la media poblacional μ.
3. La desviación estándar de las medias de muestra es $\sigma/\sqrt{n}$.

En la sección 5-6 señalamos que en ocasiones podemos aproximar una distribución de probabilidad binomial con una distribución normal. Si $np \geq 5$ y $nq \geq 5$, la variable aleatoria binomial *x* se distribuye de manera aproximadamente normal, con la media y la desviación estándar dadas por $\sigma = \sqrt{npq}$. Puesto que la distribución de probabilidad binomial trata con datos discretos y la distribución normal trata con datos continuos, aplicamos la corrección por continuidad, que debe emplearse en aproximaciones normales de distribuciones binomiales.

Finalmente, en la sección 5-7 presentamos un procedimiento para determinar si los datos muestrales parecen provenir de una población con distribución normal. Algunos de los métodos estadísticos que se estudiarán posteriormente en este libro requieren, de forma flexible, de una población que se distribuya normalmente. En estos casos es probable que lo único que se necesite sea el examen de un histograma y de los datos distantes. En otros casos se necesitarían gráficas cuantilares normales, porque es muy estricto el requisito de que la población tenga una distribución normal.

Ejercicios de repaso

1. Niveles altos de colesterol Los niveles de colesterol sérico de hombres entre 18 y 24 años de edad se distribuyen normalmente, con una media de 178.1 y una desviación estándar de 40.7. Las unidades son mg/100 mL y los datos se basan en el National Health Survey.
   a. Si se selecciona al azar a un hombre entre 18 y 24 años, calcule la probabilidad de que su nivel de colesterol sérico sea mayor que 260, valor que se considera "moderadamente alto".
   b. Si se selecciona al azar a un hombre entre 18 y 24 años, calcule la probabilidad de que su nivel de colesterol sérico esté entre 170 y 200.
   c. Si se selecciona al azar a nueve hombres entre 18 y 24 años, calcule la probabilidad de que su nivel medio de colesterol sérico esté entre 170 y 200.
   d. La Providence Health Maintenance Organization desea establecer un criterio para recomendar cambios en la dieta, si los niveles de colesterol se encuentran dentro del 3% superior. ¿Cuál es el punto de corte para los hombres de 18 a 24 años?

2. Bebés en riesgo El peso de los bebés recién nacidos en Estados Unidos se distribuye normalmente, con una media de 3420 g y una desviación estándar de 495 g (según datos de "Birth Weight and Prenatal Mortality", de Wilcox *et al*., *Journal of the American Medical Association,* vol. 273, núm. 9).
   a. Se considera que un recién nacido con un peso menor de 2200 g se encuentra en riesgo, porque la tasa de mortalidad de este grupo es al menos del 1%. ¿Qué porcentaje de recién nacidos se encuentra en la categoría "de riesgo"? Si el Chicago General Hospital tiene 900 nacimientos en un año, ¿cuántos de estos bebés se encuentran en la categoría "de riesgo"?
   b. Si redefinimos que un bebé se encuentra en riesgo si su peso al nacer está en el 2% inferior, calcule el peso que se convierte en el punto de corte que separa a los bebés en riesgo de los que no lo están.
   c. Si se seleccionan al azar 16 bebés recién nacidos, calcule la probabilidad de que su peso medio sea mayor de 3700 g.
   d. Si se seleccionan al azar 49 bebés recién nacidos, calcule la probabilidad de que su peso medio esté entre 3300 g y 3700 g.

3. Genes azules Algunas parejas poseen características genéticas que se configuran de manera que una cuarta parte de sus descendientes tienen ojos azules. Se realiza un estudio con 100 parejas en las que se sospechan dichas características; resulta que 19 de sus 100 descendientes tienen ojos azules. Suponiendo que una cuarta parte de todos los descendientes tienen ojos azules, estime la probabilidad de que, de 100 descendientes, 19 o menos tengan ojos azules. Con base en esta probabilidad, ¿parece que la tasa de un cuarto es incorrecta? ¿Por qué?

4. Estatura requerida para hombres de la Marina La Marina de Estados Unidos requiere que los hombres tengan una estatura de entre 64 y 78 pulgadas. (La National Health Survey indica que la estatura de los hombres se distribuye normalmente, con una media de 69.0 pulgadas y una desviación estándar de 2.8 pulgadas).
   a. Calcule el porcentaje de hombres que cumplen con la estatura requerida. ¿Habrá demasiados hombres a quienes se les negará la oportunidad de unirse a la Marina porque son muy bajos o muy altos?
   b. Si a usted se le designa secretario de Defensa y desea modificar el requisito de modo que sólo se rechace al 2% de los hombres más bajos y al 2% de los hombres más altos, ¿cuáles serían las nuevas estaturas mínima y máxima requeridas?
   c. Si se seleccionan 64 hombres al azar, calcule la probabilidad de que su estatura media sea mayor que 68.0 pulgadas.

5. Distribución uniforme La San Francisco Supply Company diseñó una máquina que tiene contenedores de café de modo que los contenidos se distribuyen *uniformemente*, con un mínimo de 11.8 onzas y un máximo de 12.2 onzas. Si se selecciona un contenedor al azar, calcule la probabilidad de que la cantidad de café sea

a. Menor que 12.0 onzas.
b. Entre 11.2 y 12.7 onzas.
c. Mayor que 12.2 onzas.
d. Entre 11.9 y 12.0 onzas.

6. Distribuciones muestrales

a. Se seleccionan al azar muchas muestras diferentes, de tamaño 100, de los pesos de los automóviles que se registran actualmente en Estados Unidos. ¿Qué se concluye acerca de la forma de la distribución de las medias de las distintas muestras?
b. Si los pesos de todos los automóviles que se registran en Estados Unidos tienen una desviación estándar de 512 libras, ¿cuál es la desviación estándar de las medias de muestra calculadas de muchas muestras diferentes de tamaño 100?
c. Se seleccionan al azar muchas muestras diferentes, de tamaño 1200, de la población de todos los adultos de Estados Unidos. En cada muestra se registra la proporción de personas que votaron en las últimas elecciones. ¿Qué concluye acerca de la forma de la distribución de estas proporciones de muestra?

7. Discriminación por género Cuando a varias mujeres no las contrató la Telektronics Company, se dieron a la tarea de realizar una investigación y encontraron que, entre la gran cantidad de personas que solicitaron empleo, el 30% eran mujeres. Sin embargo, las 20 personas que sí contrataron incluyen sólo dos mujeres y 18 hombres. Calcule la probabilidad de seleccionar al azar 20 personas de un grupo grande de solicitantes (30% de las cuales son mujeres) y obtener dos o menos mujeres. De acuerdo con el resultado, ¿parece que la compañía está discriminando con base en el género?

8. Prueba de normalidad Remítase a los pesos de paquetes de azúcar que se listan en el conjunto de datos 28 del Apéndice B. ¿Provienen dichos pesos de una población con distribución normal? Explique.

Ejercicios de repaso acumulativos

1. Estadísticas de movimientos oculares La lista de distancias muestreadas (en milímetros) se obtuvo con el uso de un pupilómetro para medir las distancias entre las pupilas de adultos (según datos que reunió un alumno del autor).

67	66	59	62	63	66	66	55

a. Calcule la media de las distancias en esta muestra.
b. Calcule la mediana de las distancias en esta muestra.
c. Calcule la moda de las distancias en esta muestra.
d. Calcule la desviación estándar s de esta muestra.
e. Transforme la distancia de 59 mm a una puntuación z.
f. Calcule el porcentaje real de los valores de esta muestra que excede los 59 mm.
g. Suponiendo una distribución normal, calcule los porcentajes de las distancias *poblacionales* que exceden los 59 mm. Use los valores muestrales de x y s como estimados de μ y σ.
h. ¿Qué nivel de medición (nominal, ordinal, intervalo, razón) describe este conjunto de datos?
i. Las mediciones listadas parecen redondeadas al milímetro más cercano, pero ¿las distancias exactas que no se redondearon son datos discretos o continuos?

2. Zurdos Según datos de la American Medical Association, el 10% de las personas son zurdas.

a. Si se seleccionan tres personas al azar, calcule la probabilidad de que sean zurdas.

b. Si se seleccionan tres personas al azar, calcule la probabilidad de que al menos una de ellas sea zurda.

c. ¿Por qué no podemos resolver el problema del inciso *b* a través de la aproximación normal de la distribución binomial?

d. Si se seleccionan al azar grupos de 50 personas, ¿cuál es el número medio de individuos zurdos en estos grupos?

e. Si se seleccionan al azar grupos de 50 personas, ¿cuál es la desviación estándar del número de personas zurdas en estos grupos?

f. ¿Sería infrecuente obtener ocho sujetos zurdos en un grupo seleccionado aleatoriamente de 50 personas? ¿Por qué?

Actividades de cooperación en equipo

1. Actividad fuera de clase Forme grupos de tres o cuatro estudiantes. En cada grupo, diseñe un procedimiento original para ilustrar el teorema del límite central. El objetivo principal es demostrar que cuando se seleccionan al azar muestras de una población, las medias de dichas muestras tienden a distribuirse *normalmente*, sin importar la naturaleza de la distribución poblacional. En la sección 5-5, por ejemplo, utilizamos los últimos cuatro dígitos de números del seguro social como fuente de muestras de una población de dígitos igualmente probables; procedimos a demostrar que, aun cuando la población original no tenía una distribución normal, las medias de muestras tendían a distribuirse normalmente.

2. Actividad en clase Forme grupos de tres o cuatro estudiantes. Utilice una moneda para simular nacimientos y pida que cada miembro de un grupo simule 25 nacimientos y registre el número de niñas simuladas. Combine todos los resultados del grupo y registre n = número total de nacimientos y x = número de niñas. Con los lotes de n nacimientos, calcule la media y la desviación estándar del número de niñas. ¿Es común o poco común el resultado simulado? ¿Por qué?

3. Actividad en clase Forme grupos de tres o cuatro estudiantes. Ubique los números de lotería en el conjunto de datos 26 del Apéndice B. Hay seis números que se seleccionaron aleatoriamente para cada uno de los 40 diferentes juegos de lotería. Combine los 240 números en un gran conjunto de datos y realice una prueba de normalidad. Después, calcule las 40 medias correspondientes a los 40 diferentes juegos de lotería y realice una prueba de normalidad. ¿Qué concluye? ¿Qué concepto realmente importante se ilustra en este proyecto?

4. Actividad en clase Forme grupos de tres o cuatro estudiantes. Seleccione un conjunto de datos del Apéndice B (excluya los conjuntos de datos 1, 3, 11, 12, 19 y 28, que se utilizaron como ejemplos o ejercicios en la sección 5-7). Aplique los métodos de la sección 5-7 y construya un histograma y una gráfica cuantilar normal; después, determine si el conjunto de datos parece provenir de una población distribuida normalmente.

Proyecto tecnológico

En la sección 5-3 incluimos un ejemplo sobre el diseño de automóviles. La solución en el ejemplo demostró que el 97.72% de los hombres tienen estaturas, al sentarse, menores que 38.8 pulgadas. Esa solución implicó cálculos teóricos que se basaron en el supuesto de que los hombres tienen estaturas, al sentarse, que se distribuyen normalmente, con una media de 36.0 pulgadas y una desviación estándar de 1.4 pulgadas (de acuerdo con datos de una encuesta antropométrica de Gordon, Clauser *et al*.). Este proyecto describe un método de solución diferente, que se basó en una técnica de *simulación*: usaremos una computadora o una calculadora TI-83 Plus para generar aleatoriamente 500 estaturas de hombres sentados (de una población distribuida normalmente con $\mu = 36.0$ y $\sigma = 1.4$), después calcularemos el porcentaje de los pesos simulados que sean menores que 38.8 pulgadas. A continuación se describen los procedimientos para el STATDISK, Minitab, Excel y la calculadora TI-83 Plus.

STATDISK Seleccione **Data** de la barra del menú principal, después elija la opción de **Normal Generator**. Proceda a generar 500 valores con una media de 36.0 y una desviación estándar de 1.4. (Use la opción **Format** para especificar un decimal). Después, ordene los datos con las opciones **Data**, **Sampler Editor** y luego **Format**. Con esta lista es más fácil contar el número de estaturas menores que 38.8 pulgadas. Se divide ese número entre 500 para obtener el porcentaje de las estaturas que se simulan de hombres sentados, que son menores que 38.8. Compare los resultados con el valor teórico de 97.72%, que se calculó en la sección 5-3.

Minitab Seleccione la opción **Calc**, también **Random Data**, luego **Normal**. Introduzca 500 en el número de renglones, C1 en la columna para almacenar los datos, 36.0 en el valor de la media y 1.4 en el valor de la desviación estándar. Ahora seleccione la opción **Manip**, luego **Sort** y proceda a ordenar la columna C1, con la columna ordenada y almacenada en la columna C1 y con el ordenamiento por hacer en la columna C1. Examine los valores en la columna C1 y determine el número de estaturas que se simularon que son menores que 38.8, después divida ese número entre 500 para obtener el porcentaje menor a 38.8 pulgadas. Compare los resultados con el valor teórico de 97.72% que se obtuvo en la sección 5-3.

Excel Seleccione **Tools** de la barra del menú principal, después **Data Analysis** y **Random Number Generation**. Tras hacer clic en **OK**, utilice el cuadro de diálogo para introducir uno para el número de variables y 500 para la cantidad de números aleatorios; tras esto, seleccione “normal” para el tipo de distribución. Introduzca 36.0 en la media y 1.4 en la desviación estándar. Examine los valores que se desplegaron y determine el número de estaturas que se simularon y que sean menores que 38.8, después divida ese número entre 500 para obtener el porcentaje menor a 38.8 pulgadas. Compare los resultados con el valor teórico de 97.72% que se obtuvo en la sección 5-3.

TI-83 Plus Presione **MATH**, luego **PRB**, también introduzca **randNorm** (36.0, 1.4, 500) para generar 500 valores de una población distribuida normalmente, con $\mu = 36.0$ y $\sigma = 1.4$. Presione **STO→L1** para almacenar los datos en la lista L1. Ahora presione **STAT** y luego **SortA(L1)** para ordenar los datos. Examine los datos de la lista L1 para determinar el número de estaturas simuladas que son menores que 38.8, luego hay que dividir ese número entre 500 para obtener el porcentaje menor que 38.8 pulgadas. Compare los resultados con el valor teórico de 97.72% que se obtuvo en la sección 5.3.

de los DATOS *a la* DECISIÓN

Pensamiento crítico: diseño de un asiento de automóvil

Si el asiento de un coche es muy bajo o muy alto, éste será incómodo y posiblemente peligroso. La mayoría de los asientos son ajustables, de manera que hombres y mujeres con diferentes estaturas al estar sentados pueden seleccionar una posición cómoda. Al diseñar asientos de automóviles, la altura de las rodillas de hombres y mujeres es muy importante. Los hombres tienen alturas de rodillas que se distribuyen normalmente, con una media de 22.0 pulgadas y una desviación estándar de 1.1 pulgadas; las mujeres tienen alturas de rodillas distribuidas normalmente, con una media de 20.3 pulgadas y una desviación estándar de 1.0 pulgadas (según datos de una encuesta antropométrica de Gordon, Churchill *et al.*). Calcule las alturas de rodillas máxima y mínima que incluyan al menos al 95% de todos los hombres y al menos al 95% de todas las mujeres, pero trate de calcular límites costo-beneficio tan cercanos como sea posible. Obtenga el porcentaje de hombres con alturas de rodillas entre los límites que ha determinado; después calcule el porcentaje de mujeres con alturas de rodillas entre los mismos límites. ¿Favorecen sus límites a un género, a expensas del otro? ¿Por qué no es práctico diseñar sencillamente asientos de automóvil que se ajusten a cualquiera? Si usted fuese un ingeniero de diseño de General Motors, ¿qué porcentaje de la población estaría dispuesto a excluir en su diseño de asientos para automóvil? Además de la altura de las rodillas, ¿qué otro componente de diseño importante debe tomarse en cuenta cuando se determina el rango de ajuste de los asientos de automóviles?

PROYECTO DE INTERNET

Exploración del teorema del límite central

El teorema del límite central es uno de los resultados más importantes en estadística; también puede ser uno de los más sorprendentes. De manera informal, el teorema del límite central dice que la distribución normal está en todas partes. No importa qué distribución de probabilidad subyace a un experimento, hay una distribución correspondiente de medias que tendrá una forma aproximadamente normal.

http://www.pearsoneducacion.net/triola

La mejor manera para comprender y apreciar el teorema del límite central es verlo en acción. El proyecto de Internet de este capítulo, que se encuentra en el sitio de Internet de *Estadística elemental*, le permitirá hacerlo. Se le pedirá observar, interpretar y comentar una demostración del teorema del límite central como parte de un experimento con dados. Además, será guiado a través de una búsqueda en Internet para encontrar otras demostraciones como ésta.

La estadística @ en el trabajo

"Es posible ser un periodista y no sentirse cómodo con la estadística, pero definitivamente se está limitado en lo que se puede hacer".

Joel B. Obermayer

Reportero de The News & Observer

Joel B. Obermayer escribe acerca de temas médicos y asuntos de salud para *The News & Observer*, un periódico que cubre la región este de Carolina del Norte. Realiza reportes sobre la administración de la salud, salud pública e investigaciones en centros médicos académicos, que incluyen a Duke University y University of North Carolina en Chapel Hill.

¿Qué conceptos de estadística utiliza?

Utilizo ideas como la significancia estadística, porcentajes de error y probabilidad. No necesito hacer cosas increíblemente sofisticadas, pero necesito sentirme muy cómodo con las matemáticas y con el planteamiento de preguntas acerca de ellas.

Utilizo la estadística para analizar investigaciones médicas y decidir si diferentes estudios son significativos y la forma en que escribo acerca de eso. Principalmente, necesito ser capaz de leer estadísticos y comprenderlos, más que desarrollarlos. Empleo la estadística para plantear buenas preguntas y fundamentar los argumentos que escribo. También la utilizo para decidir si alguien está tratando de darme un punto de vista positivo sobre algo que puede ser cuestionable. Por ejemplo, en una ocasión una persona de una universidad local me envió un artículo sobre cremas milagrosas que se supone que bajan de peso disolviendo las células de grasa. Pues bien, yo dudo que estas cremas funcionen. El estudio no era muy bueno tampoco. Estaban tratando de hacer aseveraciones con base en un estudio de sólo 11 personas. El investigador argumentó que 11 individuos eran suficientes para sacar buenas conclusiones empíricas sobre la salud. Eso no fue muy impresionante. Las personas tratan de manipular los medios de comunicación todo el tiempo. Los buenos estudios verificables, con buenas bases estadísticas verificables, ayudan a evitar la manipulación.

¿El uso que usted hace de la probabilidad y la estadística está aumentando, disminuyendo o permanece estable?

Está aumentando. El interés de la gente en nuevas terapias que pueden estar en la etapa de ensayo clínico se está incrementando, en parte por el énfasis en investigaciones sobre el SIDA y en la obtención de nuevos fármacos que aprobarán y recetarán pronto a los pacientes. Es más importante que nunca que un escritor médico utilice la estadística para asegurarse de que los estudios realmente prueban lo que la gente de relaciones públicas asegura que prueban.

¿Deben tener estudios de estadística los prospectos de empleados?

Es posible ser un periodista y no sentirse cómodo con la estadística, pero definitivamente se está limitado en lo que se puede hacer. Si usted escribe acerca de la eficacia de programas educativos que financia el gobierno, o si escribe acerca de los peligros de contaminantes particulares del ambiente, necesitará utilizar la estadística.

En mi campo, los editores no suelen pensar sobre la estadística en los procesos de entrevista; se preocupan más por las habilidades de escritura. El conocimiento de la estadística es más importante para lo que se puede hacer una vez que se obtiene el empleo.