1 Generalidades y estadística descriptiva

INTRODUCCIÓN

Los conceptos y métodos estadísticos no son sólo útiles sino que con frecuencia son indispensables para entender el mundo que nos rodea. Proporcionan formas de obtener ideas nuevas del comportamiento de muchos fenómenos que se presentarán en su campo de especialización escogido en ingeniería o ciencia.

La disciplina de estadística nos enseña cómo realizar juicios inteligentes y tomar decisiones informadas entre la presencia de incertidumbre y variación. Sin incertidumbre y variación, habría poca necesidad de métodos estadísticos o de profesionales en estadística. Si cada componente de un tipo particular tuviera exactamente la misma duración, si todos los resistores producidos por un fabricante tuvieran el mismo valor de resistencia, si las determinaciones del pH en muestras de suelo de un lugar particular dieran resultados idénticos, y así sucesivamente, entonces una sola observación revelaría toda la información deseada.

Una importante manifestación de variación surge en el curso de la medición de emisiones en vehículos automotores. Los requerimientos de costo y tiempo del Federal Test Procedure (FTP, por sus siglas en inglés) impiden su uso generalizado en programas de inspección de vehículos. En consecuencia, muchas agencias han creado pruebas menos costosas y más rápidas, las que se espera reproduzcan los resultados obtenidos con el FTP. De acuerdo con el artículo "Motor Vehicle Emissions Variability" (*J. of the Air and Waste Mgmt. Assoc*., 1996: 667-675), la aceptación del FTP como patrón de oro ha llevado a la creencia ampliamente difundida de que las mediciones repetidas en el mismo vehículo conducirían a resultados idénticos (o casi idénticos). Los autores del artículo aplicaron el FTP a siete vehículos caracterizados como "altos emisores". He aquí los resultados de uno de los vehículos.

HC (g/milla)	13.8	18.3	32.2	32.5
CO (g/milla)	118	149	232	236

La variación sustancial en las mediciones tanto de HC como de CO proyecta una duda considerable sobre la sabiduría convencional y hace mucho más difícil realizar evaluaciones precisas sobre niveles de emisiones.

¿Cómo se pueden utilizar técnicas estadísticas para reunir información y sacar conclusiones? Supóngase, por ejemplo, que un ingeniero de materiales inventó un recubrimiento para retardar la corrosión en tuberías de metal en circunstancias específicas. Si este recubrimiento se aplica a diferentes segmentos de la tubería, la variación de las condiciones ambientales y de los segmentos mismos producirá más corrosión sustancial en algunos segmentos que en otros. Se podría utilizar un análisis estadístico en datos de dicho experimento para decidir si la cantidad promedio de corrosión excede un límite superior especificado de alguna clase o para predecir cuánta corrosión ocurrirá en una sola pieza de tubería.

Por otra parte, supóngase que el ingeniero inventó el recubrimiento con la creencia de que será superior al recubrimiento actualmente utilizado. Se podría realizar un experimento comparativo para investigar esta cuestión aplicando el recubrimiento actual a algunos segmentos de la tubería y el nuevo a otros segmentos. Esto debe realizarse con cuidado o se obtendrá una conclusión errónea. Por ejemplo, tal vez la cantidad promedio de corrosión sea idéntica con los dos recubrimientos. Sin embargo, el recubrimiento nuevo puede ser aplicado a segmentos que tengan una resistencia superior a la corrosión y en condiciones ambientales severas en comparación con los segmentos y condiciones del recubrimiento actual. El investigador probablemente observaría entonces una diferencia entre los dos recubrimientos atribuibles no a los recubrimientos mismos, sino sólo a variaciones extrañas. La estadística ofrece no sólo métodos para analizar resultados de experimentos una vez que se han realizado sino también sugerencias sobre cómo pueden realizarse los experimentos de una manera eficiente para mitigar los efectos de variación y tener una mejor oportunidad de llegar a conclusiones correctas.

1.1 Poblaciones, muestras y procesos

Los ingenieros y científicos constantemente están expuestos a la recolección de hechos o **datos**, tanto en sus actividades profesionales como en sus actividades diarias. La disciplina de estadística proporciona métodos de organizar y resumir datos y de sacar conclusiones basadas en la información contenida en los datos.

Una investigación típicamente se enfocará en una colección bien definida de objetos que constituyen una **población** de interés. En un estudio, la población podría consistir de todas las cápsulas de gelatina de un tipo particular producidas durante un periodo específico. Otra investigación podría implicar la población compuesta de todos los individuos que recibieron una licenciatura de ingeniería durante el año académico más reciente. Cuando la información deseada está disponible para todos los objetos de la población, se tiene lo que se llama un **censo**. Las restricciones de tiempo, dinero y otros recursos escasos casi siempre hacen que un censo sea impráctico o infactible. En su lugar, se selecciona un subconjunto de la población, **una muestra**, de manera prescrita. Así pues, se podría obtener una

muestra de cojinetes de una corrida de producción particular como base para investigar si los cojinetes se ajustan a las especificaciones de fabricación, o se podría seleccionar una muestra de los graduados de ingeniería del último año para obtener retroalimentación sobre la calidad de los programas de estudio de ingeniería.

Por lo general, existe interés sólo en ciertas características de los objetos en una población: el número de grietas en la superficie de cada recubrimiento, el espesor de cada pared de cápsula, el género de un graduado de ingeniería, la edad a la cual el individuo se graduó, y así sucesivamente. Una característica puede ser categórica, tal como el género o tipo de funcionamiento defectuoso o puede ser de naturaleza numérica. En el primer caso, el *valor* de la característica es una categoría (p. ej., femenino o soldadura insuficiente), mientras que en el segundo caso, el valor es un número (p. ej., edad = 23 años o diámetro = 0.502 cm). Una **variable** es cualquier característica cuyo valor puede cambiar de un objeto a otro en la población. Inicialmente las letras minúsculas del alfabeto denotarán las variables. Algunos ejemplos incluyen:

x = marca de la calculadora de un estudiante

y = número de visitas a un sitio web particular durante un periodo específico

z = distancia de frenado de un automóvil en condiciones específicas

Se obtienen datos al observar o una sola variable o en forma simultánea dos o más variables. Un conjunto de datos **univariantes** se compone de observaciones realizadas en una sola variable. Por ejemplo, se podría determinar el tipo de transmisión automática (A) o manual (M) en cada uno de diez automóviles recientemente adquiridos en cierto concesionario y el resultado sería el siguiente conjunto de datos categóricos

M A A A M A A M A A

La siguiente muestra de duraciones (horas) de baterías D puestas en cierto uso es un conjunto de datos numéricos univariantes:

5.6 5.1 6.2 6.0 5.8 6.5 5.8 5.5

Se tienen datos **bivariantes** cuando se realizan observaciones en cada una de dos variables. El conjunto de datos podría consistir en un par (altura, peso) por cada jugador integrante del equipo de básquetbol, con la primera observación como (72, 168), la segunda como (75, 212), y así sucesivamente. Si un ingeniero determina el valor tanto de x = componente de duración y y = razón de la falla del componente, el conjunto de datos resultante es bivariante con una variable numérica y la otra categórica. Los datos **multivariantes** surgen cuando se realizan observaciones en más de una variable (por lo que bivariante es un caso especial de multivariante). Por ejemplo, un médico investigador podría determinar la presión sanguínea sistólica, la presión sanguínea diastólica y nivel de colesterol en suero de cada paciente participante en un estudio. Cada observación sería un triple de números, tal como (120, 80, 146). En muchos conjuntos de datos multivariantes, algunas variables son numéricas y otras son categóricas. Por lo tanto, el número anual dedicado al automóvil de *Consumer Reports* da valores de tales variables como tipo de vehículo (pequeño, deportivo, compacto, tamaño mediano, grande), eficiencia de consumo de combustible en la ciudad (mpg), eficiencia de consumo de combustible en carretera (mpg), tipo de tren motriz (ruedas traseras, ruedas delanteras, cuatro ruedas), etcétera.

Ramas de la estadística

Es posible que un investigador que ha recopilado datos desee resumir y describir características importantes de los mismos. Esto implica utilizar métodos de **estadística descriptiva**. Algunos de ellos son de naturaleza gráfica; la construcción de histogramas, diagramas de caja y gráficas de puntos son ejemplos primordiales. Otros métodos descriptivos implican

el cálculo de medidas numéricas, tales como medias, desviaciones estándar y coeficientes de correlación. La amplia disponibilidad de programas de computadora estadísticos han hecho que estas tareas sean más fáciles de realizar de lo que antes eran. Las computadoras son mucho más eficientes que los seres humanos para calcular y crear imágenes (¡una vez que han recibido las instrucciones apropiadas del usuario!). Esto significa que el investigador no tiene que esforzarse mucho en el "trabajo tedioso" y tendrá más tiempo para estudiar los datos y extraer mensajes importantes. A lo largo de este libro, se presentarán los datos de salida de varios paquetes tales como MINITAB, SAS, S-Plus y R. El programa R puede ser descargado sin cargo del sitio http://www.r-project.org.

Ejemplo 1.1 La tragedia que sufrió el transbordador espacial *Challenger* y sus astronautas en 1986 condujo a varios estudios para investigar las razones de la falla de la misión. La atención se enfocó de inmediato en el comportamiento de los sellos anulares del motor del cohete. He aquí datos derivados de observaciones en x = temperatura del sello anular (°F) en cada encendido de prueba o lanzamiento del motor del cohete del transbordador (*Presidential Commission on the Space Shuttle Challenger Accident*, Vol. 1, 1986: 129-131).

84	49	61	40	83	67	45	66	70	69	80	58
68	60	67	72	73	70	57	63	70	78	52	67
53	67	75	61	70	81	76	79	75	76	58	31

Sin organización, es difícil tener una idea de cuál podría ser una temperatura típica o representativa, ya sea que los valores estén muy concentrados en torno a un valor típico o bastante esparcidos, ya sea que existan brechas en los datos, qué porcentaje de los valores están en los 60, y así sucesivamente. La figura 1.1 muestra lo que se conoce como *gráfica de tallo y hojas* de los datos, así como también un *histograma*. En breve, se discutirá la construcción e interpretación de estos resúmenes gráficos; por el momento se espera que se vea cómo están distribuidos los valores de temperatura a lo largo de la escala de medición. Algunos de estos lanzamientos/encendidos fueron exitosos y otros fallaron.

```
Tallo y hojas de temperatura N = 36
Unidad de hojas = 1.0
 1    3   1
 1    3
 2    4   0
 4    4   59
 6    5   23
 9    5   788
13    6   0113
(7)   6   6777789
16    7   000023
10    7   556689
 4    8   0134
```

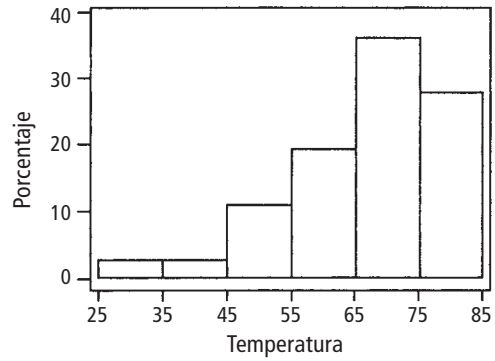


Figura 1.1 Una gráfica de tallo y hojas e histograma generados con MINITAB de los datos de temperatura de los sellos anulares.

La temperatura más baja es de 31 grados, mucho más baja que la siguiente temperatura más baja y ésta es la observación en relación con el desastre del *Challenger*. La investigación presidencial descubrió que se requerían temperaturas calientes para la operación exitosa de los sellos anulares y que 31 grados eran demasiado frío. En el capítulo 13 se presentará una relación entre temperatura y la probabilidad de un lanzamiento exitoso. ■

Después de haber obtenido una muestra de una población, un investigador con frecuencia desearía utilizar la información muestral para sacar algún tipo de conclusión (hacer una inferencia de alguna clase) con respecto a la población. Es decir, la muestra es un medio para llegar a un fin en lugar de un fin por sí misma. Las técnicas para generalizar desde una muestra hasta una población se congregan dentro de la rama de la disciplina llamada **estadística inferencial**.

Ejemplo 1.2 Las investigaciones de resistencia de materiales constituyen una rica área de aplicación de métodos estadísticos. El artículo "Effects of Aggregates and Microfillers on the Flexural Properties of Concrete" (*Magazine of Concrete Research*, 1997: 81-98) reportó sobre un estudio de propiedades de resistencia de concreto de alto desempeño obtenido con el uso de superplastificantes y ciertos aglomerantes. La resistencia a la compresión de dicho concreto previamente había sido investigada, pero no se sabía mucho sobre la resistencia a la flexión (una medida de la capacidad de resistir fallas a flexión). Los datos anexos sobre resistencia a la flexión (en megapascales, MPa, donde 1 Pa (pascal) $= 1.45 \times 10^{-4}$ lb/pulg^2) aparecieron en el artículo citado:

5.9	7.2	7.3	6.3	8.1	6.8	7.0	7.6	6.8	6.5	7.0	6.3	7.9	9.0
8.2	8.7	7.8	9.7	7.4	7.7	9.7	7.8	7.7	11.6	11.3	11.8	10.7	

Supóngase que se desea *estimar* el valor promedio de resistencia a la flexión de todas las vigas que pudieran ser fabricadas de esta manera (si se conceptualiza una población de todas esas vigas, se trata de estimar la media poblacional). Se puede demostrar que, con un alto grado de confianza, la resistencia media de la población se encuentra entre 7.48 MPa y 8.80 MPa; esto se llama *intervalo de confianza* o *estimación de intervalo*. Alternativamente, se podrían utilizar estos datos para predecir la resistencia a la flexión de una *sola* viga de este tipo. Con un alto grado de confianza, la resistencia de una sola viga excederá de 7.35 MPa; el número 7.35 se conoce como *límite de predicción inferior*. ■

El objetivo principal de este libro es presentar e ilustrar métodos de estadística inferencial que son útiles en el trabajo científico. Los tipos más importantes de procedimientos inferenciales, estimación puntual, comprobación de hipótesis y estimación por medio de intervalos de frecuencia, se introducen en los capítulos 6 a 8 y luego se utilizan escenarios más complicados en los capítulos 9 a 16. El resto de este capítulo presenta métodos de estadística descriptiva que se utilizan mucho en el desarrollo de inferencia.

Los capítulos 2 a 5 presentan material de la disciplina de probabilidad. Este material finalmente tiende un puente entre las técnicas descriptivas e inferenciales. El dominio de la probabilidad permite entender mejor cómo se desarrollan y utilizan los procedimientos inferenciales, cómo las conclusiones estadísticas pueden ser traducidas al lenguaje diario e interpretadas y cuándo y dónde pueden ocurrir errores al aplicar los métodos. La probabilidad y estadística se ocupan de cuestiones que implican poblaciones y muestras, pero lo hacen de una "manera inversa" una con respecto a la otra.

En un problema de probabilidad, se supone que las propiedades de la población estudiada son conocidas (p. ej., en una población numérica, se puede suponer una cierta distribución específica de valores de la población) y se pueden plantear y responder preguntas con respecto a una muestra tomada de una población. En un problema de estadística, el experimentador dispone de las características de una muestra y esta información le permite sacar conclusiones con respecto a la población. La relación entre las dos disciplinas se resume diciendo que la probabilidad discurre de la población a la muestra (razonamiento deductivo),

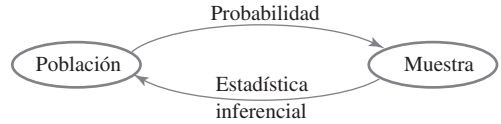


Figura 1.2 Relación entre probabilidad y estadística inferencial.

mientras que la estadística inferencial discurre de la muestra a la población (razonamiento inductivo). Esto se ilustra en la figura 1.2.

Antes de que se pueda entender lo que una muestra particular pueda decir sobre la población, primero se deberá entender la incertidumbre asociada con la toma de una muestra de una población dada. Por eso se estudia la probabilidad antes que la estadística.

Como un ejemplo del enfoque contrastante de la probabilidad y la estadística inferencial, el uso que los conductores hacen de los cinturones de seguridad manuales de regazo en carros equipados con sistemas de cinturones de hombro automáticos. (El artículo "Automobile Seat Belts: Usage Patterns in Automatic Belt Systems", *Human Factors*, 1998: 126-135, resume datos de uso.) Se podría suponer que probablemente 50% de todos los conductores de carros equipados de esta forma en cierta área metropolitana utilizan de manera regular su cinturón de regazo (una suposición sobre la población), así que se podría preguntar, "¿qué tan probable es que una muestra de 100 conductores incluirá por lo menos 70 que regularmente utilicen su cinturón de regazo?" o "¿cuántos de los conductores en una muestra de tamaño 100 se puede esperar que utilicen con regularidad su cinturón de regazo?" Por otra parte, en estadística inferencial se dispone de información sobre la muestra; por ejemplo, una muestra de 100 conductores de tales vehículos reveló que 65 utilizan con regularidad su cinturón de regazo. Se podría entonces preguntar: "¿proporciona esto evidencia sustancial para concluir que más de 50% de todos los conductores en esta área utilizan con regularidad su cinturón de regazo?" En el último escenario, se intenta utilizar la información relativa a la muestra para responder una pregunta acerca de la estructura de toda la población de la cual se seleccionó la muestra.

En el ejemplo del cinturón de regazo, la población está bien definida y concreta: todos los conductores de carros equipados de una cierta manera en un área metropolitana particular. En el ejemplo 1.1, sin embargo, una muestra de temperaturas de sello anular está disponible, pero proviene de una población que en realidad no existe. En su lugar, conviene pensar en la población como compuesta de todas las posibles mediciones de temperatura que se podrían hacer en condiciones experimentales similares. Tal población se conoce como **población conceptual** o **hipotética**. Existen varias situaciones en las cuales las preguntas encajan en el marco de referencia de la estadística inferencial al conceptualizar una población.

Estudios enumerativos contra analíticos

W. E. Deming, estadístico estadounidense muy influyente quien fue una fuerza propulsora en la revolución de calidad de Japón durante las décadas de 1950 y 1960, introdujo la distinción entre *estudios enumerativos* y *estudios analíticos*. En los primeros, el interés se enfoca en un conjunto de individuos u objetos finito, identificable y no cambiante que conforman una población. Un *marco de muestreo*, es decir, una lista de los individuos u objetos que tienen que ser muestreados, está disponible para un investigador o puede ser construida. Por ejemplo, el marco se podría componer de todas las firmas incluidas en una petición para calificar una cierta iniciativa para las boletas de votación en una elección próxima; por lo general se elige una muestra para indagar si el número de firmas *válidas* sobrepasa un valor especificado. Como otro ejemplo, el marco puede contener números de serie de todos los hornos fabricados por una compañía particular durante cierto periodo; se puede seleccionar una muestra para inferir algo sobre la duración promedio de estas unidades. El uso de métodos inferenciales presentados en este libro es razonablemente no controversial en tales escenarios (aun cuando los estadísticos continúan argumentando sobre qué métodos particulares deben ser utilizados).

Un estudio analítico se define ampliamente como uno que no es de naturaleza enumerativa. Tales estudios a menudo se realizan con el objetivo de mejorar un producto futuro al actuar sobre un proceso de una cierta clase (p. ej., recalibrar equipo o ajustar el nivel de alguna sustancia tal como la cantidad de un catalizador). A menudo se obtienen datos sólo sobre un proceso existente, uno que puede diferir en aspectos importantes del proceso futuro. No existe por lo tanto un marco de muestreo que enliste los individuos u objetos de interés. Por ejemplo, una muestra de cinco turbinas con un nuevo diseño puede ser fabricada y probada para investigar su eficiencia. Estas cinco podrían ser consideradas como una muestra de la población conceptual de todos los prototipos que podrían ser fabricados en condiciones similares, pero *no* necesariamente representativas de la población de las unidades fabricadas una vez que la producción futura esté en proceso. Los métodos para utilizar la información sobre muestras para sacar conclusiones sobre unidades de producción futuras pueden ser problemáticos. Se deberá llamar a alguien con los conocimientos necesarios en el área del diseño e ingeniería de turbinas (o de cualquier otra área pertinente) para que juzgue si tal extrapolación es sensible. Una buena exposición de estos temas se encuentra en el artículo "Assumptions for Statistical Inference", de Gerald Hahn y William Meeker (*The American Statistician*, 1993: 1-11).

Recopilación de datos

La estadística se ocupa no sólo de la organización y análisis de datos una vez que han sido recopilados sino también con el desarrollo de técnicas de recopilación de datos. Si éstos no son apropiadamente recopilados, un investigador no puede ser capaz de responder las preguntas consideradas con un razonable grado de confianza. Un problema común es que la población objetivo, aquella sobre la cual se van a sacar conclusiones, puede ser diferente de la población realmente muestreada. Por ejemplo, a los publicistas les gustaría contar con varias clases de información sobre los hábitos de ver televisión de sus clientes potenciales. La información más sistemática de esta clase proviene de colocar dispositivos de monitoreo en un pequeño número de casas a través de Estados Unidos. Se ha conjeturado que la colocación de semejantes dispositivos por sí misma modifica el comportamiento del televidente, de modo que las características de la muestra pueden ser diferentes de aquellas de la población objetivo.

Cuando la recopilación de datos implica seleccionar individuos u objetos de un marco, el método más simple para garantizar una selección representativa es tomar una *muestra aleatoria simple*. Ésta es una para la cual cualquier subconjunto particular del tamaño especificado (p. ej., una muestra de tamaño 100) tiene la misma oportunidad de ser seleccionada. Por ejemplo, si el marco se compone de 1 000 000 de números de serie, los números 1, 2, . . . , hasta 1 000 000 podrían ser anotados en trozos idénticos de papel. Después de colocarlos en una caja y mezclarlos perfectamente, se sacan uno por uno hasta que se obtenga el tamaño de muestra requisito. De manera alternativa (y mucho más preferible), se podría utilizar una tabla de números aleatorios o un generador de números aleatorios de computadora.

En ocasiones se pueden utilizar métodos de muestreo alternativos para facilitar el proceso de selección, a fin de obtener información extra o para incrementar el grado de confianza en conclusiones. Un método como ése, el *muestreo estratificado*, implica separar las unidades de la población en grupos no traslapantes y tomar una muestra de cada uno. Por ejemplo, un fabricante de reproductores de DVD podría desear información sobre la satisfacción del cliente para unidades producidas durante el año previo. Si tres modelos diferentes fueran fabricados y vendidos, se podría seleccionar una muestra distinta de cada uno de los estratos correspondientes. Esto daría información sobre los tres modelos y garantizaría que ningún modelo estuviera sobre o subrepresentado en toda la muestra.

Con frecuencia, se obtiene una muestra de "conveniencia" seleccionando individuos u objetos sin aleatorización sistemática. Por ejemplo, un conjunto de ladrillos puede ser apilado

de tal modo que sea extremadamente difícil seleccionar a los que se encuentran en el centro. Si los ladrillos localizados en la parte superior y a los lados de la pila fueran de algún modo diferentes a los demás, los datos muestrales resultantes no representarían la población. A menudo un investigador supondrá que tal muestra de conveniencia representa en forma aproximada una muestra aleatoria, en cuyo caso el repertorio de métodos inferenciales de un estadístico puede ser utilizado; sin embargo, ésta es una cuestión de criterio. La mayoría de los métodos aquí analizados se basan en una variación del muestreo aleatorio simple descrito en el capítulo 5.

Los ingenieros y científicos a menudo reúnen datos realizando alguna clase de experimento. Esto puede implicar cómo asignar varios tratamientos diferentes (tales como fertilizantes o recubrimientos anticorrosivos) a las varias unidades experimentales (parcelas o tramos de tubería). Por otra parte, un investigador puede variar sistemáticamente los niveles o categorías de ciertos factores (p. ej., presión o tipo de material aislante) y observar el efecto en alguna variable de respuesta (tal como rendimiento de un proceso de producción).

Ejemplo 1.3 Un artículo en el *New York Times* (27 de enero de 1987) reportó que el riesgo de sufrir un ataque cardiaco podría ser reducido tomando aspirina. Esta conclusión se basó en un experimento diseñado que incluía tanto un grupo de control de individuos que tomaron un placebo que tenía la apariencia de aspirina pero que se sabía era inerte y un grupo de tratamiento que tomó aspirina de acuerdo con un régimen específico. Los sujetos fueron asignados al azar a los grupos para protegerlos contra cualquier prejuicio de modo que se pudieran utilizar métodos basados en la probabilidad para analizar los datos. De los 11 034 individuos en el grupo de control, 189 subsecuentemente experimentaron ataques cardiacos, mientras que sólo 104 de los 11 037 en el grupo de aspirina sufrieron un ataque cardiaco. La tasa de incidencia de ataques cardiacos en el grupo de tratamiento fue de sólo aproximadamente la mitad de aquella en el grupo de control. Una posible explicación de este resultado es la variación de la probabilidad, que la aspirina en realidad no tiene el efecto deseado y la diferencia observada es sólo una variación típica del mismo modo que el lanzamiento al aire de dos monedas idénticas por lo general produciría diferente cantidad de águilas. No obstante, en este caso, los métodos inferenciales sugieren que la variación de la probabilidad por sí misma no puede explicar en forma adecuada la magnitud de la diferencia observada. ■

Ejemplo 1.4 Un ingeniero desea investigar los efectos tanto del tipo de adhesivo como del material conductor en la fuerza adhesiva cuando se monta un circuito integrado (CI) sobre cierto sustrato. Se consideraron dos tipos de adhesivos y dos materiales conductores. Se realizaron dos observaciones por cada combinación de tipo de adhesivo/material conductor y se obtuvieron los datos anexos.

Tipo de adhesivo	Material conductor	Fuerza de adhesión observada	Promedio
1	1	82, 77	79.5
1	2	75, 87	81.0
2	1	84, 80	82.0
2	2	78, 90	84.0

Las fuerzas adhesivas promedio resultantes se ilustran en la figura 1.3. Parece que el adhesivo tipo 2 mejora la fuerza adhesiva en comparación con el tipo 1 en aproximadamente la misma cantidad siempre que se utiliza uno de los materiales conductores, con la combinación 2, 2 como la mejor. De nuevo se pueden utilizar métodos inferenciales para juzgar si estos efectos son reales o simplemente se deben a la variación de la probabilidad.

Supóngase además que se consideran dos tiempos de curado y también dos tipos de posrecubrimientos de los circuitos integrados. Existen entonces $2 \cdot 2 \cdot 2 \cdot 2 = 16$ combinaciones de estos cuatro factores y es posible que el ingeniero no disponga de suficientes

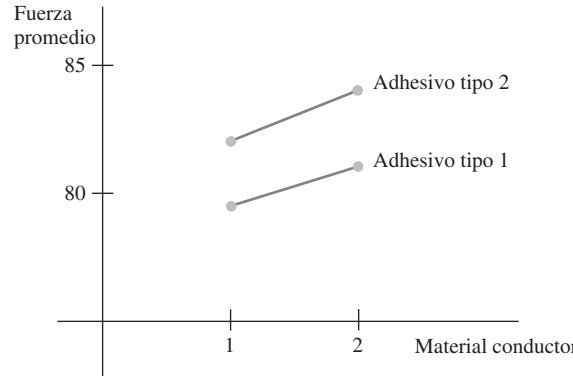


Figura 1.3 Fuerzas de adhesión promedio en el ejemplo 1.4.

recursos para hacer incluso una observación sencilla para cada una de estas combinaciones. En el capítulo 11 se verá cómo la selección cuidadosa de una fracción de estas posibilidades usualmente dará la información deseada. ■

EJERCICIOS Sección 1.1 (1-9)

1. Dé una posible muestra de tamaño 4 de cada una de las siguientes poblaciones.
- **a.** Todos los periódicos publicados en Estados Unidos.
- **b.** Todas las compañías listadas en la Bolsa de Valores de Nueva York.
- **c.** Todos los estudiantes en su colegio o universidad.
- **d.** Todas las calificaciones promedio de los estudiantes en su colegio o universidad.

2. Para cada una de las siguientes poblaciones hipotéticas, dé una muestra posible de tamaño 4.
- **a.** Todas las distancias que podrían resultar cuando usted lanza un balón de fútbol americano.
- **b.** Las longitudes de las páginas de libros publicados de aquí a 5 años.
- **c.** Todas las mediciones de intensidades posibles de terremotos (escala de Richter) que pudieran registrarse en California durante el siguiente año.
- **d.** Todos los posibles rendimientos (en gramos) de una cierta reacción química realizada en un laboratorio.

3. Considere la población compuesta de todas las computadoras de una cierta marca y modelo y enfóquese en si una computadora necesita servicio mientras se encuentra dentro de la garantía.
- **a.** Plantee varias preguntas de probabilidad con base en la selección de 100 de esas computadoras.
- **b.** ¿Qué pregunta de estadística inferencial podría ser respondida determinando el número de dichas computadoras en una muestra de tamaño 100 que requieren servicio de garantía?

4. **a.** Dé tres ejemplos diferentes de poblaciones concretas y tres ejemplos distintos de poblaciones hipotéticas.
- **b.** Por cada una de sus poblaciones concretas e hipotéticas, dé un ejemplo de una pregunta de probabilidad y un ejemplo de pregunta de estadística inferencial.

5. Muchas universidades y colegios han instituido programas de instrucción suplementaria (IS), en los cuales un facilitador regularmente se reúne con un pequeño grupo de estudiantes inscritos en el curso para promover discusiones sobre el material incluido en el curso y mejorar el dominio de la materia. Suponga que los estudiantes inscritos en un largo curso de estadística (¿de qué más?) se dividen al azar en un grupo de control que no participará en la instrucción suplementaria y en un grupo de tratamiento que sí participará. Al final del curso, se determina la calificación total de cada estudiante en el curso.
- **a.** ¿Son las calificaciones del grupo IS una muestra de una población existente? De ser así, ¿cuál es? De no ser así, ¿cuál es la población conceptual pertinente?
- **b.** ¿Cuál piensa que es la ventaja de dividir al azar a los estudiantes en los dos grupos en lugar de permitir que cada estudiante elija el grupo al que desea unirse?
- **c.** ¿Por qué los investigadores no pusieron a todos los estudiantes en el grupo de tratamiento? *Nota*: El artículo ("Supplemental Instruction: An Effective Component of Student Affairs Programming", *J. of College Student Devel.*, 1997:577-586) discute el análisis de datos de varios programas de instrucción suplementaria.

6. El sistema de la Universidad Estatal de California (CSU, por sus siglas en inglés) consta de 23 terrenos universitarios, desde la Estatal de San Diego en el sur hasta la Estatal Humboldt cerca de la frontera con Oregon. Un administrador de CSU desea hacer una inferencia sobre la distancia promedio entre la ciudad natal y sus terrenos universitarios. Describa y discuta diferentes métodos de muestreo, que pudieran ser empleados. ¿Éste sería un estudio enumerativo o un estudio analítico? Explique su razonamiento.

7. Cierta ciudad se divide naturalmente en diez distritos. ¿Cómo podría seleccionar un valuador de bienes raíces una muestra de casas unifamiliares que pudiera ser utilizada como base para desarrollar una ecuación para predecir el valor estimado a partir de características tales como antigüedad, tamaño, número de baños, distancia a la escuela más cercana y así sucesivamente? ¿El estudio es enumerativo o analítico?

8. La cantidad de flujo a través de una válvula solenoide en el sistema de control de emisiones de un automóvil es una característica importante. Se realizó un experimento para estudiar cómo la velocidad de flujo dependía de tres factores: la longitud de la armadura, la fuerza del resorte y la profundidad de la bobina. Se eligieron dos niveles diferentes (alto y bajo) de cada factor y se realizó una sola observación del flujo por cada combinación de niveles.

a. ¿De cuántas observaciones consistió el conjunto de datos resultante?

b. ¿Este estudio es enumerativo o analítico? Explique su razonamiento.

9. En un famoso experimento realizado en 1882, Michelson y Newcomb obtuvieron 66 observaciones del tiempo que requería la luz para viajar entre dos lugares en Washington, D.C. Algunas de las mediciones (codificadas en cierta manera) fueron, 31, 23, 32, 36, −2, 26, 27 y 31.

a. ¿Por qué no son idénticas estas mediciones?

b. ¿Es éste un estudio enumerativo? ¿Por qué sí o por qué no?

1.2 Métodos pictóricos y tabulares en la estadística descriptiva

La estadística descriptiva se divide en dos temas generales. En esta sección, se considera la representación de un conjunto de datos por medio de técnicas visuales. En las secciones 1.3 y 1.4, se desarrollarán algunas medidas numéricas para conjuntos de datos. Es posible que usted ya conozca muchas técnicas visuales; tablas de frecuencia, hojas de contabilidad, histogramas, gráficas de pastel, gráficas de barras, diagramas de puntos y similares. Aquí se seleccionan algunas de estas técnicas que son más útiles y pertinentes a la estadística de probabilidad e inferencial.

Notación

Alguna notación general facilitará la aplicación de métodos y fórmulas a una amplia variedad de problemas prácticos. El número de observaciones en una muestra única, es decir, el *tamaño de muestra*, a menudo será denotado por n, de modo que $n = 4$ para la muestra de universidades {Stanford, Iowa State, Wyoming, Rochester} y también para la muestra de lecturas de pH {6.3, 6.2, 5.9, 6.5}. Si se consideran dos muestras al mismo tiempo, m y n o n_1 y n_2 se pueden utilizar para denotar los números de observaciones. Por lo tanto, si {29.7, 31.6, 30.9} y {28.7, 29.5, 29.4, 30.3} son lecturas de eficiencia térmica de dos tipos diferentes de motores diesel, entonces $m = 3$ y $n = 4$.

Dado un conjunto de datos compuesto de n observaciones de alguna variable x, entonces $x_1, x_2, x_3, \ldots, x_n$ denotarán las observaciones individuales. El subíndice no guarda ninguna relación con la magnitud de una observación particular. Por lo tanto, x_1 en general no será la observación más pequeña del conjunto, ni x_n será la más grande. En muchas aplicaciones, x_1 será la primera observación realizada por el experimentador, x_2 la segunda, y así sucesivamente. La observación i-ésima del conjunto de datos será denotada por x_i.

Gráficas de tallos y hojas

Considérese un conjunto de datos numéricos $x_1, x_2, \ldots, x_n$ para el cual x_i se compone de por lo menos dos dígitos. Una forma rápida de obtener la representación visual informativa del conjunto de datos es construir una *gráfica de tallos y hojas*.

Pasos para construir una gráfica de tallos y hojas

1. Seleccione uno o más de los primeros dígitos para los valores de tallo. Los segundos dígitos se convierten en hojas.
2. Enumere los posibles valores de tallos en una columna vertical.
3. Anote la hoja para cada observación junto al valor de tallo.
4. Indique las unidades para tallos y hojas en algún lugar de la gráfica.

Si el conjunto de datos se compone de calificaciones de exámenes, cada uno entre 0 y 100, la calificación de 83 tendría un tallo de 8 y una hoja de 3. Para un conjunto de datos de eficiencias de consumo de combustible de automóviles (mpg), todas entre 8.1 y 47.8, se podrían utilizar como el tallo, así que 32.6 tendría entonces una hoja de 2.6. En general, se recomienda una gráfica basada en tallos entre 5 y 20.

Ejemplo 1.5 El consumo de alcohol por parte de estudiantes universitarios preocupa no sólo a la comunidad académica sino también, a causa de consecuencias potenciales de salud y seguridad, a la sociedad en su conjunto. El artículo ("Health and Behavioral Consequences of Binge Drinking in College", *J. of the Amer. Med. Assoc.*, 1994: 1672-1677) presentó un amplio estudio sobre el consumo excesivo de alcohol en universidades a través de Estados Unidos. Un episodio de parranda se definió como cinco o más tragos en fila para varones y cuatro o más para mujeres. La figura 1.4 muestra una gráfica de tallo y hojas de 140 valores de $x =$ porcentaje de edades de los estudiantes de licenciatura bebedores. (Estos valores no aparecieron en el artículo citado, pero la gráfica concuerda con una gráfica de los datos que sí lo hicieron.)

```
0 | 4
1 | 1345678889
2 | 1223456666777889999
3 | 01122333445556666777778888999999
4 | 11122222334444556666667778888899 9
5 | 00111222233455666667777888899
6 | 01111244455666778
```

Tallo: dígitos de diez cifras
Hojas: dígitos de una cifra

Figura 1.4 Gráfica de tallo y hojas de porcentajes de bebedores en cada una de 140 universidades.

La primera hoja de la fila 2 del tallo es 1, la cual dice que 21% de los estudiantes de una de las universidades de la muestra eran bebedores. Sin la identificación de los dígitos en los tallos y los dígitos en las hojas, no se sabría si la observación correspondiente al tallo 2, hoja 1 debería leerse como 21%, 2.1% o 0.21 por ciento.

Cuando se crea una imagen a mano, la ordenación de las hojas de la más pequeña a la más grande en cada línea puede ser tediosa. Esta ordenación contribuye poco si no se dispone de información adicional. Supóngase que las observaciones hubieran sido puestas en lista en orden alfabético por nombre de la escuela, como

16% 33% 64% 37% 31% . . .

Entonces la colocación de estos valores en la gráfica en este orden haría que la fila 1 del tallo tuviera 6 como su primera hoja y el principio de la fila 3 del tallo sería

3 | 371 . . .

La gráfica sugiere que un valor típico o representativo se encuentra en la fila 4 del tallo, tal vez en el rango medio de 40%. Las observaciones no aparecen muy concentradas en torno a este valor típico, como sería el caso si todos los valores estuvieran entre 20 y 49%. Esta gráfica se eleva a una sola cresta a medida que desciende, y luego declina; no hay brechas en la gráfica. La forma de la gráfica no es perfectamente simétrica, pero en su lugar parece alargarse un poco más en la dirección de las hojas bajas que en la dirección de las hojas altas. Por último, no existen observaciones que se alejen inusualmente del grueso de los datos (ningunos *valores apartados*), como sería el caso si uno de los valores de 26% hubiera sido de 86%. La característica más sobresaliente de estos datos es que, en la mayoría de las universidades de la muestra, por lo menos una cuarta parte de los estudiantes son bebedores. El problema de beber en exceso en las universidades es mucho más extenso de lo que muchos hubieran sospechado. ■

Una gráfica de tallos y hojas da información sobre los siguientes aspectos de los datos:

- Identificación de un valor típico o representativo.
- Grado de dispersión en torno al valor típico.
- Presencia de brechas en los datos.
- Grado de simetría en la distribución de los valores.
- Número y localización de crestas.
- Presencia de valores afuera de la gráfica.

Ejemplo 1.6 La figura 1.5 presenta gráficas de tallos y hojas de una muestra aleatoria de longitudes de campos de golf (yardas) designados por *Golf Magazine* como los más desafiantes en Estados Unidos. Entre la muestra de 40 campos, el más corto es de 6 433 yardas de largo y el más largo es de 7 280 yardas. Las longitudes parecen estar distribuidas de una manera aproximadamente uniforme dentro del rango de valores presentes en la muestra. Obsérvese que la selección de tallo en este caso de un solo dígito (6 ó 7) o de tres (643, . . . , 728) produciría una gráfica no informativa, primero a causa de pocos tallos y segundo a causa de demasiados.

Los programas de computadora de estadística en general no producen gráficas con tallos de dígitos múltiples. La gráfica MINITAB que aparece en la figura 1.5(b) resulta de *truncar* cada observación al borrar los dígitos uno.

64	35	64	33	70			
65	26	27	06	83			
66	05	94	14				
67	90	70	00	98	70	45	13
68	90	70	73	50			
69	00	27	36	04			
70	51	05	11	40	50	22	
71	31	69	68	05	13	65	
72	80	09					

Tallo: dígitos de miles y cientos de cifras
Hojas: dígitos de decenas de cifras y una cifra

a)

```
Tallo y hojas de yardaje  N = 40
Unidad de hojas = 10
          4          64  3367
          8          65  0228
         11          66  019
         18          67  0147799
        (4)          68  5779
         18          69  0023
         14          70  012455
          8          71  013666
          2          72  08
```

b)

Figura 1.5 Gráficas de tallo y hojas de yardajes de campos de golf: a) hojas de dos dígitos; b) gráfica generada por MINITAB con las hojas de un dígito truncadas. ■

Gráficas de puntos

Una gráfica de puntos es un resumen atractivo de datos numéricos cuando el conjunto de datos es razonablemente pequeño o existen pocos valores de datos distintos. Cada observación está representada por un punto sobre la ubicación correspondiente en una escala de medición horizontal. Cuando un valor ocurre más de una vez, existe un punto por cada ocurrencia y estos puntos se apilan verticalmente. Como con la gráfica de tallos y hojas, una gráfica de puntos da información sobre la localización, dispersión, extremos y brechas.

Ejemplo 1.7 La figura 1.6 muestra una gráfica de puntos para los datos de temperatura de los sellos anulares introducidos en el ejemplo 1.1 en la sección previa. Un valor de temperatura representativo es uno que se encuentra entre la mitad de los 60 (°F) y existe poca dispersión en torno al centro. Los datos se alargan más en el extremo inferior que en el superior y la observación más pequeña, 31, apenas puede ser descrita como valor extremo.

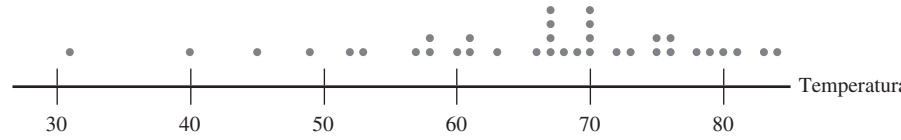


Figura 1.6 Gráfica de puntos de los datos de temperatura de los sellos anulares (°F). ■

Si el conjunto de datos del ejemplo 1.7 hubieran consistido en 50 o 100 observaciones de temperatura, cada una registrada a un décimo de grado, habría sido muy tedioso construir una gráfica de puntos. La técnica siguiente es muy adecuada a situaciones como esas.

Histogramas

Algunos datos numéricos se obtienen contando para determinar el valor de una variable (el número de citatorios de tráfico que una persona recibió durante el año pasado, el número de personas que solicitan empleo durante un periodo particular), mientras que otros datos se obtienen tomando mediciones (peso de un individuo, tiempo de reacción a un estímulo particular). La prescripción para trazar un histograma es en general diferente en estos dos casos.

DEFINICIÓN

Una variable numérica es **discreta** si su conjunto de valores posibles es finito o se puede enumerar en una sucesión infinita (una en la cual existe un primer número, un segundo número, y así sucesivamente). Una variable numérica es **continua** si sus valores posibles abarcan un intervalo completo sobre la línea de números.

Una variable discreta x casi siempre resulta de contar, en cuyo caso posibles valores son 0, 1, 2, 3, . . . o algún subconjunto de estos enteros. De la toma de mediciones surgen variables continuas. Por ejemplo, si x es el pH de una sustancia química, entonces en teoría x podría ser cualquier número entre 0 y 14: 7.0, 7.03, 7.032 y así sucesivamente. Desde luego, en la práctica existen limitaciones en el grado de precisión de cualquier instrumento de medición, por lo que es posible que no se pueda determinar el pH, el tiempo de reacción, la altura y la concentración con un número arbitrariamente grande de decimales. Sin embargo, desde el punto de vista de crear modelos matemáticos de distribuciones de datos, conviene imaginar un conjunto completo continuo de valores posibles.

Considérense datos compuestos de observaciones de una variable discreta x. La **frecuencia** de cualquier valor x particular es el número de veces que ocurre un valor en el conjunto de datos. La **frecuencia relativa** de un valor es la fracción o proporción de veces que ocurre el valor:

$$\text{frecuencia relativa de un valor} = \frac{\text{número de veces que ocurre el valor}}{\text{número de observaciones en el conjunto de datos}}$$

Supóngase, por ejemplo, que el conjunto de datos se compone de 200 observaciones de x = el número de cursos que un estudiante está tomando en este semestre. Si 70 de estos valores x es 3, entonces

$$\text{frecuencia del valor 3 de } x\text{: } 70$$

$$\text{frecuencia relativa del valor 3 de } x\text{: } \frac{70}{200} = 0.35$$

Si se multiplica una frecuencia relativa por 100 se obtiene un porcentaje en el ejemplo de cursos universitarios, 35% de los estudiantes de la muestra están tomando tres cursos. Las

frecuencias relativas, o porcentajes, por lo general interesan más que las frecuencias mismas. En teoría, las frecuencias relativas deberán sumar 1, pero en la práctica la suma puede diferir un poco de 1 por el redondeo. Una **distribución de frecuencia** es una tabla de las frecuencias o de las frecuencias relativas, o de ambas.

> Construcción de un histograma para datos discretos
>
> En primer lugar, se determina la frecuencia y la frecuencia relativa de cada valor x. Luego se marcan los valores x posibles en una escala horizontal. Sobre cada valor, se traza un rectángulo cuya altura es la frecuencia relativa (o alternativamente, la frecuencia) de dicho valor.

Esta construcción garantiza que el *área* de cada rectángulo es proporcional a la frecuencia relativa del valor. Por lo tanto, si las frecuencias relativas de $x = 1$ y $x = 5$ son 0.35 y 0.07, respectivamente, entonces el área del rectángulo sobre 1 es cinco veces el área del rectángulo sobre 5.

Ejemplo 1.8 ¿Qué tan inusual es un juego de béisbol sin hit o de un hit en las ligas mayores y cuán frecuentemente un equipo pega más de 10, 15 o incluso 20 hits? La tabla 1.1 es una distribución de frecuencia del número de hits por equipo por juego de todos los juegos de nueve episodios que se jugaron entre 1989 y 1993.

Tabla 1.1 Distribución de frecuencia de hits en juegos de nueve episodios

Hits/juego	Número de juegos	Frecuencia relativa	Hits/juego	Número de juegos	Frecuencia relativa
0	20	0.0010	14	569	0.0294
1	72	0.0037	15	393	0.0203
2	209	0.0108	16	253	0.0131
3	527	0.0272	17	171	0.0088
4	1048	0.0541	18	97	0.0050
5	1457	0.0752	19	53	0.0027
6	1988	0.1026	20	31	0.0016
7	2256	0.1164	21	19	0.0010
8	2403	0.1240	22	13	0.0007
9	2256	0.1164	23	5	0.0003
10	1967	0.1015	24	1	0.0001
11	1509	0.0779	25	0	0.0000
12	1230	0.0635	26	1	0.0001
13	834	0.0430	27	1	0.0001
				19 383	1.0005

El histograma correspondiente en la figura 1.7 se eleva suavemente hasta una sola cresta y luego declina. El histograma se extiende un poco más hacia la derecha (hacia valores grandes) que hacia la izquierda, un poco "asimétrico positivo".

O con la información tabulada o con el histograma mismo, se puede determinar lo siguiente:

$$\begin{aligned}\text{proporción de juegos a lo sumo de dos hits} &= \begin{matrix}\text{frecuencia}\\ \text{relativa}\\ \text{de } x = 0\end{matrix} + \begin{matrix}\text{frecuencia}\\ \text{relativa}\\ \text{de } x = 1\end{matrix} + \begin{matrix}\text{frecuencia}\\ \text{relativa}\\ \text{de } x = 2\end{matrix}\\ &= 0.0010 + 0.0037 + 0.0108 = 0.0155\end{aligned}$$

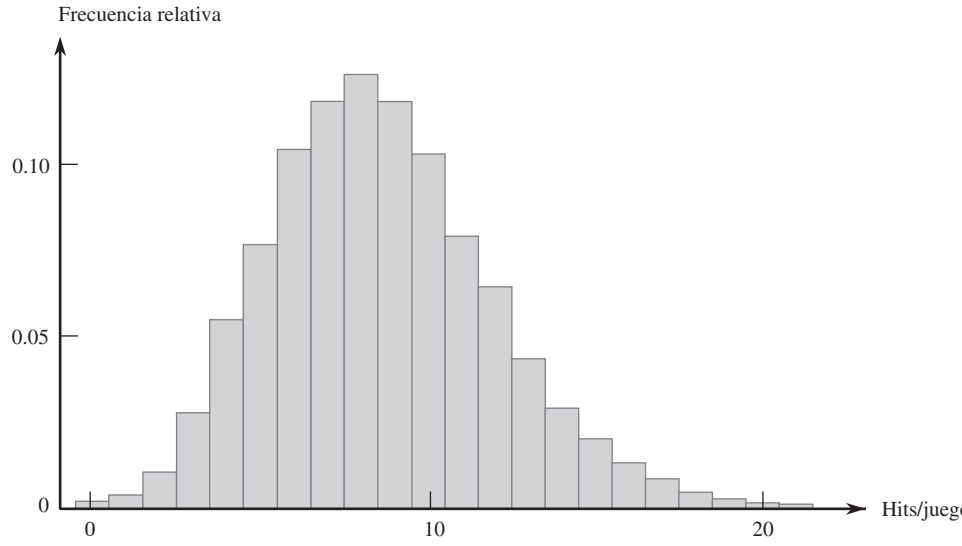


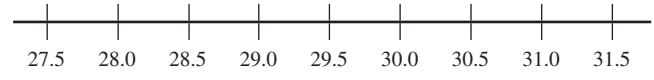
Figura 1.7 Histograma de número de hits por juego de nueve episodios.

Asimismo,

$$\text{proporción de juegos con entre 5 y 10 hits (inclusive)} = 0.0752 + 0.1026 + \cdots + 0.1015 = 0.6361$$

Esto es, aproximadamente 64% de todos estos juegos fueron de entre 5 y 10 hits (inclusive). ■

La construcción de un histograma para datos continuos (mediciones) implica subdividir el eje de medición en un número adecuado de **intervalos de clase** o **clases**, de tal suerte que cada observación quede contenida en exactamente una clase. Supóngase, por ejemplo, que se hacen 50 observaciones de x = eficiencia de consumo de combustible de un automóvil (mpg), la más pequeña de las cuales es 27.8 y la más grande 31.4. Entonces se podrían utilizar los límites de clase 27.5, 28.0, 28.5, . . . , y 31.5 como se muestra a continuación:



Una dificultad potencial es que de vez en cuando una observación está en un límite de clase así que por consiguiente no cae en exactamente un intervalo, por ejemplo, 29.0. Una forma de habérselas con este problema es utilizar límites como 27.55, 28.05, . . . , 31.55. La adición de centésimas a los límites de clase evita que las observaciones queden en los límites resultantes. Otro método es utilizar las clases 27.5–<28.0, 28.0–<28.5, . . . , 31.0–<31.5. En ese caso 29.0 queda en la clase 29.0–<29.5 y no en la clase 28.5–<29.0. En otras palabras, con esta convención, una observación que queda en el límite se coloca en el intervalo a la *derecha* del mismo. Así es como MINITAB construye un histograma.

Construcción de un histograma para datos continuos: anchos de clase iguales

Se determina la frecuencia y la frecuencia relativa de cada clase. Se marcan los límites de clase sobre un eje de medición horizontal. Sobre cada intervalo de clase, se traza un rectángulo cuya altura es la frecuencia relativa correspondiente (o frecuencia).

Ejemplo 1.9 Las compañías eléctricas requieren información sobre el consumo de los clientes para obtener pronósticos precisos de demandas. Investigadores de Wisconsin Power and Light determinaron el consumo de energía (BTU) durante un periodo particular con una muestra de 90 hogares calentados con gas. Se calculó un valor de consumo promedio como sigue:

$$\text{consumo ajustado} = \frac{\text{consumo}}{\text{(clima, en grados días)(área de casa)}}$$

Esto dio por resultado los datos anexos (una parte del conjunto de datos guardados FURNACE.MTW disponible en MINITAB, el cual se ordenó desde el valor más pequeño al más grande).

2.97	4.00	5.20	5.56	5.94	5.98	6.35	6.62	6.72	6.78
6.80	6.85	6.94	7.15	7.16	7.23	7.29	7.62	7.62	7.69
7.73	7.87	7.93	8.00	8.26	8.29	8.37	8.47	8.54	8.58
8.61	8.67	8.69	8.81	9.07	9.27	9.37	9.43	9.52	9.58
9.60	9.76	9.82	9.83	9.83	9.84	9.96	10.04	10.21	10.28
10.28	10.30	10.35	10.36	10.40	10.49	10.50	10.64	10.95	11.09
11.12	11.21	11.29	11.43	11.62	11.70	11.70	12.16	12.19	12.28
12.31	12.62	12.69	12.71	12.91	12.92	13.11	13.38	13.42	13.43
13.47	13.60	13.96	14.24	14.35	15.12	15.24	16.06	16.90	18.26

Se permite que MINITAB seleccione los intervalos de clase. La característica del histograma en la figura 1.8 que más llama la atención es su parecido a una curva en forma de campana (y por consiguiente simétrico), con el punto de simetría aproximadamente en 10.

Frecuencia de clase	1–<3 1	3–<5 1	5–<7 11	7–<9 21	9–<11 25	11–<13 17	13–<15 9	15–<17 4	17–<19 1
Frecuencia relativa	0.011	0.011	0.122	0.233	0.278	0.189	0.100	0.044	0.011

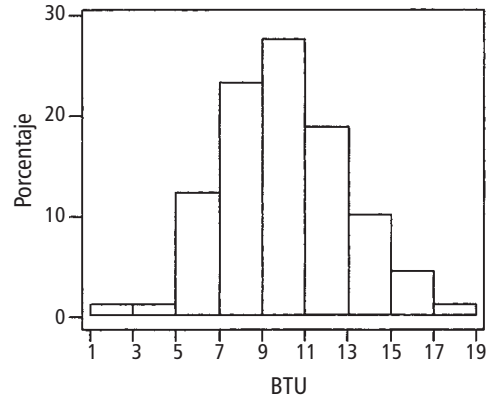


Figura 1.8 Histograma de los datos de consumo de energía del ejemplo 1.9.

De acuerdo con el histograma,

$$\begin{pmatrix}\text{proporción de}\\ \text{observaciones}\\ \text{menor que 9}\end{pmatrix} \approx 0.01 + 0.01 + 0.12 + 0.23 = 0.37 \ \left(\text{valor exacto} = \frac{34}{90} = 0.378\right)$$

La frecuencia relativa para la clase 9-<11 es aproximadamente 0.27, así que se estima que en forma aproximada la mitad de ésta, o 0.135, queda entre 9 y 10. Por lo tanto

$$\begin{matrix}\text{proporción de observaciones} \\ \text{menores que 10}\end{matrix} \approx 0.37 + 0.135 = 0.505 \text{ (poco más de 50\%)}$$

El valor exacto de esta proporción es 47/90 = 0.522 ■

No existen reglas inviolables en cuanto al número de clases o la selección de las mismas. Entre 5 y 20 serán satisfactorias para la mayoría de los conjuntos de datos. En general, mientras más grande es el número de observaciones en un conjunto de datos, más clases deberán ser utilizadas. Una razonable regla empírica es

$$\text{número de clases} \approx \sqrt{\text{número de observaciones}}$$

Es posible que las clases de ancho-igual no sean una opción sensible si un conjunto de datos "se alarga" hacia un lado o el otro. La figura 1.9 muestra una curva de puntos de dicho conjunto de datos. Con un pequeño número de clases de ancho-igual casi todas las observaciones quedan en exactamente una o dos de las clases. Si se utiliza un gran número de clases de ancho-igual las frecuencias de muchas clases será cero. Una buena opción es utilizar algunos intervalos más anchos cerca de las observaciones extremas y más angostos en la región de alta concentración.

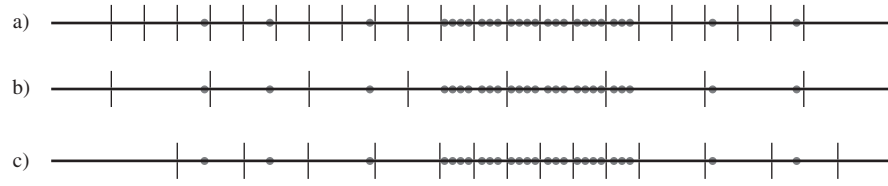


Figura 1.9 Selección de intervalos de clase para un conjunto "alargado" de puntos: a) intervalos angostos de ancho igual; b) intervalos amplios de ancho igual; c) intervalos de anchos diferentes.

Construcción de un histograma para datos continuos: anchos de clase desiguales

Después de determinar las frecuencias y las frecuencias relativas, se calcula la altura de cada rectángulo con la fórmula

$$\text{altura del rectángulo} = \frac{\text{frecuencia relativa de la clase}}{\text{ancho de clase}}$$

Las alturas del rectángulo resultante en general se conocen como *densidades* y la escala vertical es la **escala de densidades**. Esta prescripción también funcionará cuando los anchos de clase son iguales.

Ejemplo 1.10 La corrosión del acero de refuerzo es un problema serio en estructuras de concreto localizadas en ambientes afectados por condiciones climáticas severas. Por esa razón, los investigadores han estado estudiando el uso de barras de refuerzo hechas de un material compuesto. Se realizó un estudio para desarrollar indicaciones para adherir barras de refuerzo reforzadas con fibra de vidrio a concreto ("Design Recommendations for Bond of GFRP Rebars to Concrete", *J. of Structural Engr.*, 1996: 247-254). Considérense las siguientes 48 observaciones de fuerza adhesiva medida:

11.5	12.1	9.9	9.3	7.8	6.2	6.6	7.0	13.4	17.1	9.3	5.6
5.7	5.4	5.2	5.1	4.9	10.7	15.2	8.5	4.2	4.0	3.9	3.8
3.6	3.4	20.6	25.5	13.8	12.6	13.1	8.9	8.2	10.7	14.2	7.6
5.2	5.5	5.1	5.0	5.2	4.8	4.1	3.8	3.7	3.6	3.6	3.6

Frecuencia de clase	2–<4 9	4–<6 15	6–<8 5	8–<12 9	12–<20 8	20–<30 2
Frecuencia relativa	0.1875	0.3125	0.1042	0.1875	0.1667	0.0417
Densidad	0.094	0.156	0.052	0.047	0.021	0.004

El histograma resultante aparece en la figura 1.10. La cola derecha o superior se alarga mucho más que la izquierda o inferior, un sustancial alejamiento de la simetría.

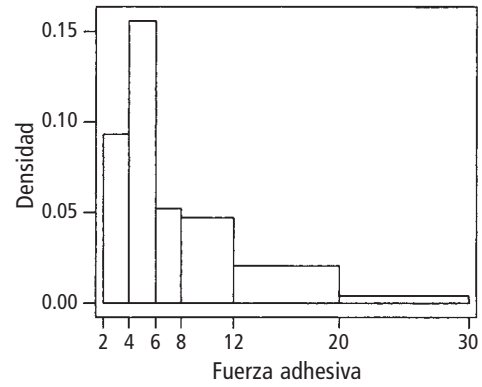


Figura 1.10 Un histograma de densidad generado por MINITAB de los datos de fuerza adhesiva del ejemplo 1.10. ■

Cuando los anchos de clase son desiguales, si no se utiliza una escala de densidad se obtendrá una gráfica con áreas distorsionadas. Con anchos de clase iguales, el divisor es el mismo en cada cálculo de densidad y la aritmética adicional simplemente implica reescalar el eje vertical (es decir, el histograma que utiliza frecuencia relativa y el que utiliza densidad tendrán exactamente la misma apariencia). Un histograma de densidad tiene una propiedad interesante. Si se multiplican ambos miembros de la fórmula para densidad por el ancho de clase se obtiene

$$\begin{aligned}\text{frecuencia relativa} &= (\text{ancho de clase})(\text{densidad})\\ &= (\text{ancho del rectángulo})(\text{altura del rectángulo}) = \text{área del rectángulo}\end{aligned}$$

Es decir, *el área de cada rectángulo es la frecuencia relativa de la clase correspondiente*. Además, como la suma de frecuencias relativas debe ser 1, *el área total de todos los rectángulos en un histograma de densidad es* 1. Siempre es posible trazar un histograma de modo que el área sea igual a la frecuencia relativa (esto es cierto también para un histograma de datos discretos), simplemente se utiliza la escala de densidad. Esta propiedad desempeñará un importante papel al crear modelos de distribución en el capítulo 4.

Formas de histograma

Los histogramas se presentan en varias formas. Un histograma **unimodal** es el que se eleva a una sola cresta y luego declina. Uno **bimodal** tiene dos crestas diferentes. Puede ocurrir bimodalidad cuando el conjunto de datos se compone de observaciones de dos clases bastante diferentes de individuos u objetos. Por ejemplo, considérese un gran conjunto de datos compuesto de tiempos de manejo de automóviles que viajan entre San Luis Obispo, California

y Monterey, California (sin contar el tiempo utilizado para ver puntos de interés, comer, etc.). Este histograma mostraría dos crestas, una para los carros que toman la ruta interior (aproximadamente 2.5 horas) y otra para los carros que viajan a lo largo de la costa (3.5-4 horas). Sin embargo, la bimodalidad no se presenta automáticamente en dichas situaciones. Sólo si los dos histogramas distintos están "muy alejados" en forma relativa con respecto a sus esparcimientos la bimodalidad ocurrirá en el histograma de datos combinados. Por consiguiente un conjunto de datos grande compuesto de estaturas de estudiantes universitarios no producirá un histograma bimodal porque la altura típica de hombres de aproximadamente 69 pulgadas no está demasiado por encima de la altura típica de mujeres de aproximadamente 64-65 pulgadas. Se dice que un histograma con más de dos crestas es **multimodal**. Por supuesto, el número de crestas dependerá de la selección de intervalos de clase, en particular, con un pequeño número de observaciones. Mientras más grande es el número de clases, es más probable que se manifieste bimodalidad o multimodalidad.

Un histograma es **simétrico** si la mitad izquierda es una imagen de espejo de la mitad derecha. Un histograma bimodal es **positivamente asimétrico** si la cola derecha o superior se alarga en comparación con la cola izquierda o inferior y **negativamente asimétrico** si el alargamiento es hacia la izquierda. La figura 1.11 muestra histogramas "alisados" obtenidos superponiendo una curva alisada sobre los rectángulos, que ilustran varias posibilidades.

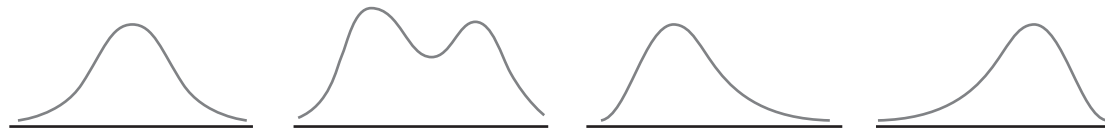


Figura 1.11 Histogramas alisados: a) unimodal simétrico; b) bimodal; c) positivamente asimétrico y d) negativamente asimétrico.

Datos cualitativos

Tanto una distribución de frecuencia y un histograma pueden ser construidos cuando el conjunto de datos es de naturaleza *cualitativa* (categórico). En algunos casos, habrá un ordenamiento natural de las clases, por ejemplo, estudiantes de primer año, segundo, tercero, cuarto y graduados, mientras que en otros casos el orden será arbitrario, por ejemplo, católico, judío, protestante, etc. Con esos datos categóricos, los intervalos sobre los que se construyen rectángulos deberán ser de igual ancho.

Ejemplo 1.11 El Public Policy Institute of California realizó una encuesta telefónica de 2501 residentes adultos en California durante abril de 2006 para indagar qué pensaban sobre varios aspectos de la educación pública K-12. Una pregunta fue "En general, ¿cómo calificaría la calidad de las escuelas públicas de su vecindario hoy en día? La tabla 1.2 muestra las frecuencias y las frecuencias relativas y la figura 1.12 muestra el histograma correspondiente (gráfica de barras).

Tabla 1.2 Distribución de frecuencia de calificaciones escolares

Calificación	Frecuencia	Frecuencia relativa
A	478	0.191
B	893	0.357
C	680	0.272
D	178	0.071
F	100	0.040
Desconocida	172	0.069
	2501	1.000

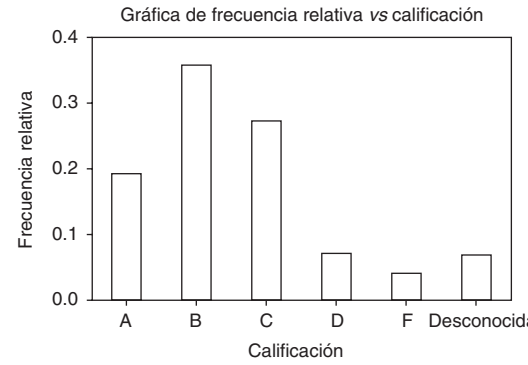


Figura 1.12 Histograma de calificaciones de las escuelas obtenido con MINITAB.

Más de la mitad de los encuestados otorgaron una calificación A o B y sólo un poco más de 10% otorgó una calificación D o F. Los porcentajes de padres de niños que asisten a escuelas públicas fueron un poco más favorables para las escuelas: 24, 40, 24, 6, 4 y 2 por ciento. ■

Datos multivariantes

Los datos multivariantes en general son más difíciles de describir en forma visual. Varios métodos para hacerlo aparecen más adelante en el libro, notablemente en gráficas de puntos de datos numéricos bivariantes.

EJERCICIOS Sección 1.2 (10-32)

10. Considere los datos de resistencia de las vigas del ejemplo 1.2.

- **a.** Construya una gráfica de tallos y hojas de los datos. ¿Cuál parece ser el valor de resistencia representativo? ¿Parecen estar las observaciones altamente concentradas en torno al valor representativo o algo dispersas?
- **b.** ¿Parece ser la gráfica razonablemente simétrica en torno a un valor representativo o describiría su forma de otra manera?
- **c.** ¿Parece haber algunos valores de resistencia extremos?
- **d.** ¿Qué proporción de las observaciones de resistencia en esta muestra exceden de 10 MPa?

11. Cada calificación en el siguiente lote de calificaciones de exámenes se encuentra en los 60, 70, 80 o 90. Una gráfica de tallos y hojas con sólo los cuatro tallos 6, 7, 8 y 9 no describiría detalladamente la distribución de calificaciones. En tales situaciones, es deseable utilizar tallos repetidos. En este caso se repetiría el tallo 6 dos veces, utilizando 6L para las calificaciones en los 60 bajos (hojas 0, 1, 2, 3 y 4) y 6H para las calificaciones en los 60 altos (hojas 5, 6, 7, 8 y 9). Asimismo, los demás tallos pueden ser repetidos dos veces para obtener una gráfica de ocho filas. Construya la gráfica para las calificaciones dadas. ¿Qué característica de los datos es resaltada por esta gráfica?

74	89	80	93	64	67	72	70	66	85	89	81	81
71	74	82	85	63	72	81	81	95	84	81	80	70
69	66	60	83	85	98	84	68	90	82	69	72	87
88												

12. Los valores de densidad relativa anexos de varios tipos de madera utilizados en la construcción aparecieron en el artículo ("Bolted Connection Design Values Based on European Yield Model", *J. of Structural Engr*., 1993: 2169-2186):

0.31	0.35	0.36	0.36	0.37	0.38	0.40	0.40	0.40
0.41	0.41	0.42	0.42	0.42	0.42	0.42	0.43	0.44
0.45	0.46	0.46	0.47	0.48	0.48	0.48	0.51	0.54
0.54	0.55	0.58	0.62	0.66	0.66	0.67	0.68	0.75

Construya una gráfica de tallos y hojas con tallos repetidos (véase el ejercicio previo) y comente sobre cualquier característica interesante de la gráfica.

13. Las propiedades mecánicas permisibles para el diseño estructural de vehículos aeroespaciales metálicos requieren un método aprobado para analizar estadísticamente datos de prueba empíricos. El artículo ("Establishing Mechanical Property Allowables for Metals", *J. of Testing and Evaluation*, 1998: 293-299) utilizó los datos anexos sobre resistencia a la tensión última (lb/pulg^2) como base para abordar las dificultades que se presentan en el desarrollo de dicho método.

122.2	124.2	124.3	125.6	126.3	126.5	126.5	127.2	127.3
127.5	127.9	128.6	128.8	129.0	129.2	129.4	129.6	130.2
130.4	130.8	131.3	131.4	131.4	131.5	131.6	131.6	131.8
131.8	132.3	132.4	132.4	132.5	132.5	132.5	132.5	132.6
132.7	132.9	133.0	133.1	133.1	133.1	133.1	133.2	133.2
133.2	133.3	133.3	133.5	133.5	133.5	133.8	133.9	134.0
134.0	134.0	134.0	134.1	134.2	134.3	134.4	134.4	134.6

134.7	134.7	134.7	134.8	134.8	134.8	134.9	134.9	135.2
135.2	135.2	135.3	135.3	135.4	135.5	135.5	135.6	135.6
135.7	135.8	135.8	135.8	135.8	135.8	135.9	135.9	135.9
135.9	136.0	136.0	136.1	136.2	136.2	136.3	136.4	136.4
136.6	136.8	136.9	136.9	137.0	137.1	137.2	137.6	137.6
137.8	137.8	137.8	137.9	137.9	138.2	138.2	138.3	138.3
138.4	138.4	138.4	138.5	138.5	138.6	138.7	138.7	139.0
139.1	139.5	139.6	139.8	139.8	140.0	140.0	140.7	140.7
140.9	140.9	141.2	141.4	141.5	141.6	142.9	143.4	143.5
143.6	143.8	143.8	143.9	144.1	144.5	144.5	147.7	147.7

a. Construya una gráfica de tallos y hojas de los datos eliminando (truncando) los dígitos de décimos y luego repitiendo cada valor de tallo cinco veces (una vez para las hojas 1 y 2, una segunda vez para las hojas 3 y 4, etc.). ¿Por qué es relativamente fácil identificar un valor de resistencia representativo?

b. Construya un histograma utilizando clases de ancho igual con la primera clase que tiene un límite inferior de 122 y un límite superior de 124. Enseguida comente sobre cualquier característica interesante del histograma.

14. El conjunto de datos adjunto se compone de observaciones del flujo de una regadera (l/min) para una muestra de $n = 129$ casas en Perth, Australia ("An Application of Bayes Methodology to the Analysis of Diary Records in a Water Use Study", *J. Amer. Stat. Assoc*., 1987: 705-711):

4.6	12.3	7.1	7.0	4.0	9.2	6.7	6.9	11.5	5.1
11.2	10.5	14.3	8.0	8.8	6.4	5.1	5.6	9.6	7.5
7.5	6.2	5.8	2.3	3.4	10.4	9.8	6.6	3.7	6.4
8.3	6.5	7.6	9.3	9.2	7.3	5.0	6.3	13.8	6.2
5.4	4.8	7.5	6.0	6.9	10.8	7.5	6.6	5.0	3.3
7.6	3.9	11.9	2.2	15.0	7.2	6.1	15.3	18.9	7.2
5.4	5.5	4.3	9.0	12.7	11.3	7.4	5.0	3.5	8.2
8.4	7.3	10.3	11.9	6.0	5.6	9.5	9.3	10.4	9.7
5.1	6.7	10.2	6.2	8.4	7.0	4.8	5.6	10.5	14.6
10.8	15.5	7.5	6.4	3.4	5.5	6.6	5.9	15.0	9.6
7.8	7.0	6.9	4.1	3.6	11.9	3.7	5.7	6.8	11.3
9.3	9.6	10.4	9.3	6.9	9.8	9.1	10.6	4.5	6.2
8.3	3.2	4.9	5.0	6.0	8.2	6.3	3.8	6.0	

a. Construya una gráfica de tallos y hojas de los datos.

b. ¿Cuál es una velocidad de flujo o gasto típico o representativo?

c. ¿Parece estar la gráfica altamente concentrada o dispersa?

d. ¿Es la distribución de valores razonablemente simétrica? Si no, ¿cómo describiría el alejamiento de la simetría?

e. ¿Describiría cualquier observación como alejada del resto de los datos (un valor extremo)?

15. Un artículo de *Consumer Reports* sobre crema de cacahuate (septiembre de 1990) reportó las siguientes calificaciones para varias marcas:

Creamy	56	44	62	36	39	53	50	65	45	40
	56	68	41	30	40	50	56	30	22	
Crunchy	62	53	75	42	47	40	34	62	52	
	50	34	42	36	75	80	47	56	62	

Construya una gráfica de tallos y hojas *comparativa* y ponga una lista de tallos a la mitad de la página y luego coloque las hojas "creamy" a la derecha y las "crunchy" a la izquierda. Describa las similitudes y diferencias de los dos tipos.

16. El artículo citado en el ejemplo 1.2 también dio las observaciones de resistencia adjuntas para los cilindros:

6.1	5.8	7.8	7.1	7.2	9.2	6.6	8.3	7.0	8.3
7.8	8.1	7.4	8.5	8.9	9.8	9.7	14.1	12.6	11.2

a. Construya una gráfica de tallos y hojas comparativa (véase el ejercicio previo) de los datos de la viga y el cilindro y luego responda las preguntas en las partes b)-d) del ejercicio 10 para las observaciones de los cilindros.

b. ¿En qué formas son similares los dos lados de la gráfica? ¿Existen algunas diferencias obvias entre las observaciones de la viga y las observaciones del cilindro?

c. Construya una gráfica de puntos de los datos del cilindro.

17. Transductores de temperatura de cierto tipo se envían en lotes de 50. Se seleccionó una muestra de 60 lotes y se determinó el número de transductores en cada lote que no cumplen con las especificaciones de diseño y se obtuvieron los datos siguientes:

2	1	2	4	0	1	3	2	0	5	3	3	1	3	2	4	7	0	2	3
0	4	2	1	3	1	1	3	4	1	2	3	2	2	8	4	5	1	3	1
5	0	2	3	2	1	0	6	4	2	1	6	0	3	3	3	6	1	2	3

a. Determine las frecuencias y las frecuencias relativas de los valores observados de $x =$ número de transductores en un lote que no cumple con las especificaciones.

b. ¿Qué proporción de lotes muestreados tienen a lo sumo cinco transductores que no cumplen con las especificaciones? ¿Qué proporción tiene menos de cinco? ¿Qué proporción tienen por lo menos cinco unidades que no cumplen con las especificaciones?

c. Trace un histograma de los datos que utilizan la frecuencia relativa en la escala vertical y comente sus características.

18. En un estudio de productividad de autores ("Lotka's Test", *Collection Mgmt*., 1982: 111-118), se clasificó a un gran número de autores de artículos de acuerdo con el número de artículos que publicaron durante cierto periodo. Los resultados se presentaron en la distribución de frecuencia adjunta:

Número de artículos	1	2	3	4	5	6	7	8
Frecuencia	784	204	127	50	33	28	19	19
Número de artículos	9	10	11	12	13	14	15	16	17
Frecuencia	6	7	6	7	4	4	5	3	3

a. Construya un histograma correspondiente a esta distribución de frecuencia. ¿Cuál es la característica más interesante de la forma de la distribución?

b. ¿Qué proporción de estos autores publicó por lo menos cinco artículos? ¿Por lo menos diez artículos? ¿Más de diez artículos?

c. Suponga que los cinco 15, los tres 6 y los tres 17 se agruparon en una sola categoría mostrada como "≥ 15". ¿Podría trazar un histograma? Explique.

d. Suponga que los valores 15, 16 y 17 se enlistan por separado y se combinan en la categoría 15-17 con frecuencia 11. ¿Sería capaz de trazar un histograma? Explique.

19. Se determinó el número de partículas contaminadas en una oblea de silicio antes de cierto proceso de enjuague por cada oblea en una muestra de tamaño 100 y se obtuvieron las siguientes frecuencias:

Número de partículas	0	1	2	3	4	5	6	7
Frecuencia	1	2	3	12	11	15	18	10

Número de partículas	8	9	10	11	12	13	14
Frecuencia	12	4	5	3	1	2	1

a. ¿Qué proporción de las obleas muestreadas tuvieron por lo menos una partícula? ¿Por lo menos cinco partículas?

b. ¿Qué proporción de las obleas muestreadas tuvieron entre cinco y diez partículas, inclusive? ¿Estrictamente entre cinco y diez partículas?

c. Trace un histograma con la frecuencia relativa en el eje vertical. ¿Cómo describiría la forma del histograma?

20. El artículo ("Determination of Most Representative Subdivision", *J. of Energy Engr.*, 1993: 43-55) dio datos sobre varias características de subdivisiones que podrían ser utilizados para decidir si se suministra energía eléctrica con líneas elevadas o líneas subterráneas. He aquí los valores de la variable x = longitud total de calles dentro de una subdivisión:

1280	5320	4390	2100	1240	3060	4770
1050	360	3330	3380	340	1000	960
1320	530	3350	540	3870	1250	2400
960	1120	2120	450	2250	2320	2400
3150	5700	5220	500	1850	2460	5850
2700	2730	1670	100	5770	3150	1890
510	240	396	1419	2109		

a. Construya una gráfica de hojas y tallos con las milésimas como el tallo y las centésimas como las hojas y comente sobre algunas características de la gráfica.

b. Construya un histograma con los límites de clase, 0, 1000, 2000, 3000, 4000, 5000 y 6000. ¿Qué proporción de subdivisiones tienen una longitud total menor que 2000? ¿Entre 2000 y 4000? ¿Cómo describiría la forma del histograma?

21. El artículo citado en el ejercicio 20 también da los siguientes valores de las variables y = número de calles cerradas y z = número de intersecciones:

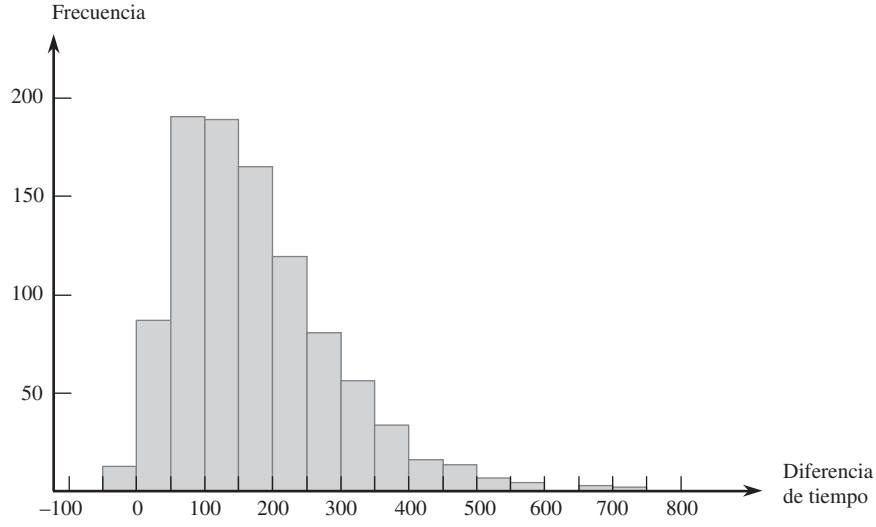
y	1	0	1	0	0	2	0	1	1	1	2	1	0	0	1	1	0	1	1
z	1	8	6	1	1	5	3	0	0	4	4	0	0	1	2	1	4	0	4
y	1	1	0	0	0	1	1	2	0	1	2	2	1	1	0	2	1	1	0
z	0	3	0	1	1	0	1	3	2	4	6	6	0	1	1	8	3	3	5
y	1	5	0	3	0	1	1	0	0										
z	0	5	2	3	1	0	0	0	3										

a. Construya un histograma con los datos y. ¿Qué proporción de estas subdivisiones no tenía calles cerradas? ¿Por lo menos una calle cerrada?

b. Construya un histograma con los datos z. ¿Qué proporción de estas subdivisiones tenía cuando mucho cinco intersecciones? ¿Menos de cinco intersecciones?

22. ¿Cómo varía la velocidad de un corredor en el recorrido del curso de un maratón (una distancia de 42.195 km)? Considere determinar tanto el tiempo de recorrido de los primeros 5 km y el tiempo de recorrido entre los 35 y 40 km, y luego reste el primer tiempo del segundo. Un valor positivo de esta diferencia corresponde a un corredor que corre más lento hacia el final de la carrera. El histograma adjunto está basado en tiempos de corredores que participaron en varios maratones japoneses ("Factors Affecting Runners' Maratón Performance", *Chance,* otoño de 1993: 24-30).

Histograma del ejercicio 22



¿Cuáles son algunas características interesantes de este histograma? ¿Cuál es un valor de diferencia típico? ¿Aproximadamente qué proporción de los competidores corren la última distancia más rápido que la primera?

23. En un estudio de ruptura de la urdimbre durante el tejido de telas (*Technometrics*, 1982: 63), se sometieron a prueba 100 muestras de hilo. Se determinó el número de ciclos de esfuerzo hasta ruptura para cada muestra de hilo y se obtuvieron los datos siguientes:

86	146	251	653	98	249	400	292	131	169
175	176	76	264	15	364	195	262	88	264
157	220	42	321	180	198	38	20	61	121
282	224	149	180	325	250	196	90	229	166
38	337	65	151	341	40	40	135	597	246
211	180	93	315	353	571	124	279	81	186
497	182	423	185	229	400	338	290	398	71
246	185	188	568	55	55	61	244	20	284
393	396	203	829	239	236	286	194	277	143
198	264	105	203	124	137	135	350	193	188

a. Construya un histograma de frecuencia relativa basado en los intervalos de clase 0-<100, 100-<200, . . . y comente sobre las características del histograma.

b. Construya un histograma basado en los siguientes intervalos de clase: 0-<50, 50-<100, 100-<150, 150-<200, 200-<300, 300-<400, 400-<500, 500-<600 y 600-<900.

c. Si las especificaciones de tejido requieren una resistencia a la ruptura de por lo menos 100 ciclos, ¿qué proporción de los especímenes de hilos en esta muestra sería considerada satisfactoria?

24. El conjunto de datos adjuntos consiste en observaciones de resistencia al esfuerzo cortante (lb) de soldaduras de puntos ultrasónicas aplicadas en un cierto tipo de lámina alclad. Construya un histograma de frecuencia relativa basado en diez clases de ancho igual con límites 4000, 4200, . . . [El histograma concordará con el que aparece en ("Comparison of Properties of Joints Prepared by Ultrasonic Welding and Other Means", *J. of Aircraft*, 1983: 552-556).] Comente sobre sus características.

5434	4948	4521	4570	4990	5702	5241
5112	5015	4659	4806	4637	5670	4381
4820	5043	4886	4599	5288	5299	4848
5378	5260	5055	5828	5218	4859	4780
5027	5008	4609	4772	5133	5095	4618
4848	5089	5518	5333	5164	5342	5069
4755	4925	5001	4803	4951	5679	5256
5207	5621	4918	5138	4786	4500	5461
5049	4974	4592	4173	5296	4965	5170
4740	5173	4568	5653	5078	4900	4968
5248	5245	4723	5275	5419	5205	4452
5227	5555	5388	5498	4681	5076	4774
4931	4493	5309	5582	4308	4823	4417
5364	5640	5069	5188	5764	5273	5042
5189	4986					

25. Una transformación de valores de datos por medio de alguna función matemática, tal como $\sqrt{x}$ o $1/x$ a menudo produce un conjunto de números que tienen "mejores" propiedades estadísticas que los datos originales. En particular, puede ser posible encontrar una función para la cual el histograma de valores transformados es más simétrico (o, incluso mejor, más parecido a una curva en forma de campana) que los datos originales. Por ejemplo, el artículo ("Time Lapse Cinematographic Analysis of Beryllium-Lung Fibroblast Interactions", *Environ. Research*, 1983: 34-43) reportó los resultados de experimentos diseñados para estudiar el comportamiento de ciertas células individuales que habían estado expuestas a berilio. Una importante característica de dichas células individuales es su tiempo de interdivisión (IDT, por sus siglas en inglés). Se determinaron tiempos de interdivisión de un gran número de células tanto en condiciones expuestas (tratamiento) como no expuestas (control). Los autores del artículo utilizaron una transformación logarítmica, es decir, valor transformado = log(valor original). Considere los siguientes tiempos de interdivisión representativos.

IDT	$\log_{10}$(IDT)	IDT	$\log_{10}$(IDT)	IDT	$\log_{10}$(IDT)
28.1	1.45	60.1	1.78	21.0	1.32
31.2	1.49	23.7	1.37	22.3	1.35
13.7	1.14	18.6	1.27	15.5	1.19
46.0	1.66	21.4	1.33	36.3	1.56
25.8	1.41	26.6	1.42	19.1	1.28
16.8	1.23	26.2	1.42	38.4	1.58
34.8	1.54	32.0	1.51	72.8	1.86
62.3	1.79	43.5	1.64	48.9	1.69
28.0	1.45	17.4	1.24	21.4	1.33
17.9	1.25	38.8	1.59	20.7	1.32
19.5	1.29	30.6	1.49	57.3	1.76
21.1	1.32	55.6	1.75	40.9	1.61
31.9	1.50	25.5	1.41		
28.9	1.46	52.1	1.72		

Use los intervalos de clase 10–<20, 20–<30, . . . para construir un histograma de los datos originales. Use los intervalos 1.1–<1.2, 1.2–<1.3, . . . para hacer lo mismo con los datos transformados. ¿Cuál es el efecto de la transformación?

26. En la actualidad se está utilizando la difracción retrodispersada de electrones en el estudio de fenómenos de fractura. La siguiente información sobre ángulo de desorientación (grados) se extrajo del artículo ("Observations on the Faceted Initiation Site in the Dwell-Fatigue Tested Ti-6242 Alloy: Crystallographic Orientation and Size Effects", *Metallurgical and Materials Trans*., 2006: 1507-1518).

Clase:	0–<5	5–<10	10–<15	15–<20
Frec. rel.:	0.177	0.166	0.175	0.136
Clase:	20–<30	30–<40	40–<60	60–<90
Frec. rel.:	0.194	0.078	0.044	0.030

a. ¿Es verdad que más de 50% de los ángulos muestreados son más pequeños que 15°, como se afirma en el artículo?

b. ¿Qué proporción de los ángulos muestreados son por lo menos de 30°?

c. ¿Aproximadamente qué proporción de los ángulos son de entre 10° y 25°?

d. Construya un histograma y comente sobre cualquier característica interesante.

27. El artículo ("Study on the Life Distribution of Microdrills", *J. of Engr. Manufacture*, 2002: 301-305) reportó las siguientes observaciones, listadas en orden creciente sobre la duración de brocas (número de agujeros que una broca fresa antes de que se rompa) cuando se fresaron agujeros en una cierta aleación de latón.

11	14	20	23	31	36	39	44	47	50
59	61	65	67	68	71	74	76	78	79
81	84	85	89	91	93	96	99	101	104
105	105	112	118	123	136	139	141	148	158
161	168	184	206	248	263	289	322	388	513

a. ¿Por qué una distribución de frecuencia no puede estar basada en los intervalos de clase 0-50, 50-100, 100-150 y así sucesivamente?

b. Construya una distribución de frecuencia e histograma de los datos con los límites de clase 0, 50, 100, . . . y luego comente sobre las características interesantes.

c. Construya una distribución de frecuencia e histograma de los logaritmos naturales de las observaciones de duración y comente sobre características interesantes.

d. ¿Qué proporción de las observaciones de duración en esta muestra son menores que 100? ¿Qué proporción de las observaciones son de por lo menos 200?

28. Las mediciones humanas constituyen una rica área de aplicación de métodos estadísticos. El artículo ("A Longitudinal Study of the Development of Elementary School Children's Private Speech", *Merrill-Palmer Q.*, 1990: 443-463) reportó sobre un estudio de niños que hablan solos (conversación a solas). Se pensaba que la conservación a solas tenía que ver con el IQ, porque se supone que éste mide la madurez mental y se sabía que la conservación a solas disminuye conforme los estudiantes avanzan a través de los años de la escuela primaria. El estudio incluyó 33 estudiantes cuyas calificaciones de IQ de primer año se dan a continuación:

82	96	99	102	103	103	106	107	108	108	108	108
109	110	110	111	113	113	113	113	115	115	118	118
119	121	122	122	127	132	136	140	146			

Describa los datos y comente sobre cualquier característica importante.

29. Considere los siguientes datos sobre el tipo de problemas de salud (J = hinchazón de las articulaciones, F = fatiga, B = dolor de espalda, M = debilidad muscular, C = tos, N = nariz suelta/irritación, O = otro) que aquejan a los plantadores de árboles. Obtenga las frecuencias y las frecuencias relativas de las diversas categorías y trace un histograma. (Los datos son consistentes con los porcentajes dados en el artículo ("Physiological Effects of Work Stress and Pesticide Exposure in Tree Planting de British Columbia Silviculture Workers", *Ergonomics*, 1993: 951-961.)

O	O	N	J	C	F	B	B	F	O	J	O	O	M
O	F	F	O	O	N	O	N	J	F	J	B	O	C
J	O	J	J	F	N	O	B	M	O	J	M	O	B
O	F	J	O	O	B	N	C	O	O	O	M	B	F
J	O	F	N										

30. Un **diagrama de Pareto** es una variación de un histograma de datos categóricos producidos por un estudio de control de calidad. Cada categoría representa un tipo diferente de no conformidad del producto o problema de producción. Las categorías se ordenaron de modo que la categoría con la frecuencia más grande aparezca a la extrema izquierda, luego la categoría con la segunda frecuencia más grande, y así sucesivamente. Suponga que se obtiene la siguiente información sobre no conformidades en paquetes de circuito: componentes averiados, 126; componentes incorrectos, 210; soldadura insuficiente, 67; soldadura excesiva, 54; componente faltante, 131. Construya un diagrama de Pareto.

31. La **frecuencia acumulativa** y la frecuencia relativa acumulativa de un intervalo de clase particular son la suma de frecuencias y frecuencias relativas, respectivamente, del intervalo y todos los intervalos que quedan debajo de él. Si, por ejemplo, existen cuatro intervalos con frecuencias 9, 16, 13 y 12, entonces las frecuencias acumulativas son 9, 25, 38 y 50 y las frecuencias relativas acumulativas son 0.18, 0.50, 0.76 y 1.00 Calcule las frecuencias acumulativas y las frecuencias relativas de los datos del ejercicio 24.

32. La carga de incendio (MJ/m^2) es la energía calorífica que podría ser liberada por metro cuadrado de área de piso por la combustión del contenido y la estructura misma. El artículo ("Fire Loads in Office Buildings", *J. of Structural Engr.*, 1997: 365-368) dio los siguientes porcentajes acumulativos (tomados de una gráfica) de cargas de fuego en una muestra de 388 cuartos:

Valor	0	150	300	450	600
% acumulativo	0	19.3	37.6	62.7	77.5
Valor	750	900	1050	1200	1350
% acumulativo	87.2	93.8	95.7	98.6	99.1
Valor	1500	1650	1800	1950	
% acumulativo	99.5	99.6	99.8	100.0	

a. Construya un histograma de frecuencia relativa y comente sobre características interesantes.

b. ¿Qué proporción de cargas de fuego es menor que 600? ¿Por lo menos de 1200?

c. ¿Qué proporción de las cargas está entre 600 y 1200?

1.3 Medidas de localización

Los resúmenes visuales de datos son herramientas excelentes para obtener impresiones y percepciones preliminares. Un análisis de datos más formal a menudo requiere el cálculo e interpretación de medidas resumidas numéricas. Es decir, de los datos se trata de extraer varios números resumidos, números que podrían servir para caracterizar el conjunto de datos

y comunicar algunas de sus características prominentes. El interés principal se concentrará en los datos numéricos; al final de la sección aparecen algunos comentarios con respecto a datos categóricos.

Supóngase, entonces, que el conjunto de datos es de la forma $x_1, x_2, \ldots, x_n$, donde cada x_i es un número. ¿Qué características del conjunto de números son de mayor interés y merecen énfasis? Una importante característica de un conjunto de números es su localización y en particular su centro. Esta sección presenta métodos para describir la localización de un conjunto de datos; en la sección 1.4 se regresará a los métodos para medir la variabilidad en un conjunto de números.

La media

Para un conjunto dado de números $x_1, x_2, \ldots, x_n$, la medida más conocida y útil del centro es la *media* o promedio aritmético del conjunto. Como casi siempre se pensará que los números x_i constituyen una muestra, a menudo se hará referencia al promedio aritmético como la *media muestral* y se la denotará por $\bar{x}$.

DEFINICIÓN

La **media muestral** $\bar{x}$ de las observaciones $x_1, x_2, \ldots, x_n$ está dada por

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

El numerador de $\bar{x}$ se escribe más informalmente como $\sum x_i$, donde la suma incluye todas las observaciones muestrales.

Para reportar $\bar{x}$, se recomienda utilizar una precisión decimal de un dígito más que la precisión de los números x_i. Por consiguiente las observaciones son distancias de detención con $x_1 = 125$, $x_2 = 131$ y así sucesivamente, se podría tener $\bar{x} = 127.3$ pies.

Ejemplo 1.12 El agrietamiento de hierro y acero provocado por corrosión producida por esfuerzo cáustico ha sido estudiado debido a las fallas que se presentan alrededor de los remaches en calderas de acero y fallas de rotores de turbinas de vapor. Considérense las observaciones adjuntas de x = longitud de agrietamiento (μm) derivadas de pruebas de corrosión con esfuerzo constante en probetas de barras pulidas sometidas a tensión durante un periodo fijo. (Los datos concuerdan con un histograma y cantidades resumidas tomadas del artículo "On the Role of Phosphorus in the Caustic Stress Corrosion Cracking of Low Alloy Steels", *Corrosion Science*, 1989: 53-68.)

$x_1 = 16.1$ $x_2 = 9.6$ $x_3 = 24.9$ $x_4 = 20.4$ $x_5 = 12.7$ $x_6 = 21.2$ $x_7 = 30.2$
$x_8 = 25.8$ $x_9 = 18.5$ $x_{10} = 10.3$ $x_{11} = 25.3$ $x_{12} = 14.0$ $x_{13} = 27.1$ $x_{14} = 45.0$
$x_{15} = 23.3$ $x_{16} = 24.2$ $x_{17} = 14.6$ $x_{18} = 8.9$ $x_{19} = 32.4$ $x_{20} = 11.8$ $x_{21} = 28.5$

La figura 1.13 muestra una gráfica de tallo y hojas de los datos; una longitud de agrietamiento en los 20 bajos parece ser "típica".

0H	96	89				
1L	27	03	40	46	18	
1H	61	85				
2L	49	04	12	33	42	Tallo: dígitos de decenas
2H	58	53	71	85		Hojas: dígitos de unidades y decenas
3L	02	24				
3H						
4L						
4H	50					

Figura 1.13 Gráfica de tallo y hojas de los datos de la longitud de agrietamiento.

Con $\sum x_i = 444.8$, la media muestral es

$$\bar{x} = \frac{444.8}{21} = 21.18$$

un valor consistente conforme a la información dada por la gráfica de tallo y hojas. ■

Una interpretación física de $\bar{x}$ demuestra cómo mide la ubicación (centro) de una muestra. Se traza y gradúa un eje de medición horizontal y luego se representa cada observación muestral por una pesa de 1 lb colocada en el punto correspondiente sobre el eje. El único punto en el cual se puede colocar un punto de apoyo para equilibrar el sistema de pesas es el punto correspondiente al valor de $\bar{x}$ (véase la figura 1.14).

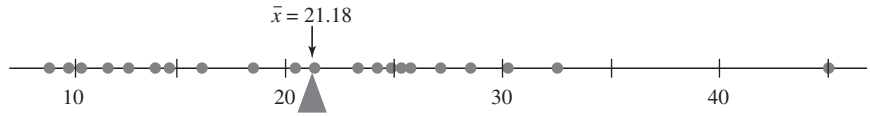


Figura 1.14 La media como punto de equilibrio de un sistema de pesas.

Así como $\bar{x}$ representa el valor promedio de las observaciones incluidas en una muestra, se puede calcular el promedio de todos los valores incluidos en la población. Este promedio se llama **media de la población** y está denotada por la letra griega μ. Cuando existen N valores en la población (una población finita), entonces μ = (suma de los N valores de población)/N. En los capítulos 3 y 4, se dará una definición más general de μ que se aplica tanto a poblaciones finitas y (conceptualmente) infinitas. Así como $\bar{x}$ es una medida interesante e importante de la ubicación de la muestra, μ es una interesante e importante característica (con frecuencia la más importante) de una población. En los capítulos de inferencia estadística, se presentarán métodos basados en la media muestral para sacar conclusiones con respecto a una media de población. Por ejemplo, se podría utilizar la media muestral $\bar{x} = 21.18$ calculada en el ejemplo 1.12 como una *estimación puntual* (un solo número que es la "mejor" conjetura) de μ = la longitud de agrietamiento promedio verdadera de todas las probetas tratadas como se describe.

La media sufre de una deficiencia que la hace ser una medida inapropiada del centro en algunas circunstancias: su valor puede ser afectado en gran medida por la presencia de incluso un solo valor extremo (una observación inusualmente grande o pequeña). En el ejemplo 1.12, el valor $x_{14} = 45.0$ es obviamente un valor extremo. Sin esta observación, $\bar{x} = 399.8/20 = 19.99$; el valor extremo incrementa la media en más de 1 μm. Si la observación de 45.0 μm fuera reemplazada por el valor catastrófico de 295.0 μm, un valor realmente extremo, entonces $\bar{x} = 694.8/21 = 33.09$, ¡el cual es más grande que todos excepto una de las observaciones!

Una muestra de ingresos a menudo produce algunos valores apartados (unos cuantos afortunados que gana cantidades astronómicas) y el uso del ingreso promedio como medida de ubicación con frecuencia será engañoso. Tales ejemplos sugieren que se busca una medida que sea menos sensible a los valores apartados que $\bar{x}$ y momentáneamente se propondrá una. Sin embargo, aunque $\bar{x}$ sí tiene este defecto potencial, sigue siendo la medida más ampliamente utilizada, en gran medida porque existen muchas poblaciones para las cuales un valor extremo en la muestra sería altamente improbable. Cuando se muestrea una población como esa (una población normal o en forma de campana es el ejemplo más importante), la media muestral tenderá a ser estable y bastante representativa de la muestra.

La mediana

La palabra *mediana* es sinónimo de "medio" y la mediana muestral es en realidad el valor medio una vez que se ordenan las observaciones de la más pequeña a la más grande. Cuando las observaciones están denotadas por $x_1, \ldots, x_n$, se utilizará el símbolo $\tilde{x}$ para representar la mediana muestral.

DEFINICIÓN

La **mediana muestral** se obtiene ordenando primero las n observaciones de la más pequeña a la más grande (con cualesquiera valores repetidos incluidos de modo que cada observación muestral aparezca en la lista ordenada). Entonces,

$$\tilde{x} = \begin{cases} \text{El valor medio único si } n \text{ es impar} & = \left(\dfrac{n+1}{2}\right)^{n\text{-ésimo}} \text{ valor ordenado} \\ \text{El promedio de los dos valores medios si } n \text{ es par} & = \text{promedio de } \left(\dfrac{n}{2}\right)^{n\text{-ésimo}} \text{ y } \left(\dfrac{n}{2}+1\right)^{n\text{-ésimo}} \text{ valores ordenados} \end{cases}$$

Ejemplo 1.13 El riesgo de desarrollar deficiencia de hierro es especialmente alto durante el embarazo. El problema con la detección de tal deficiencia es que algunos métodos para determinar el estado del hierro pueden ser afectados por el estado de gravidez mismo. Considérense las siguientes observaciones ordenadas de concentración de receptores de transferrina de una muestra de mujeres con evidencia de laboratorio de anemia por deficiencia de hierro evidente ("Serum Transferrin Receptor for the Detection of Iron Deficiency in Pregnancy", *Amer. J. of Clinical Nutrition*, 1991: 1077-1081):

7.6 8.3 9.3 9.4 9.4 9.7 10.4 11.5 11.9 15.2 16.2 20.4

Como $n = 12$ es par, el $n/2 =$ los valores sexto y séptimo ordenados deben ser promediados:

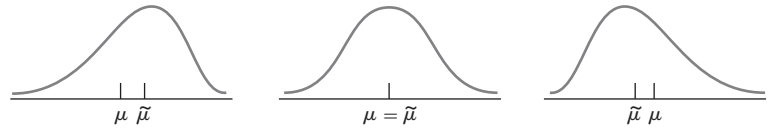
$$\tilde{x} = \frac{9.7 + 10.4}{2} = 10.05$$

Note que si la observación más grande, 20.4, no hubiera aparecido en la muestra, la mediana muestral resultante de las $n = 11$ observaciones habría sido el valor medio 9.7 [el $(n + 1)/2$ = sexto valor ordenado]. La media muestral es $\bar{x} = \sum x_i/n = 139.3/12 = 11.61$, la cual es un tanto más grande que la mediana debido a los valores apartados 15.2, 16.2 y 20.4. ■

Los datos del ejemplo 1.13 ilustran una importante propiedad de $\tilde{x}$ en contraste con $\bar{x}$. La mediana muestral es muy insensible a los valores apartados. Si, por ejemplo, las dos x_i más grandes se incrementan desde 16.2 y 20.4 hasta 26.2 y 30.4, respectivamente, $\tilde{x}$ no se vería afectada. Por lo tanto, en el tratamiento de valores apartados, $\bar{x}$ y $\tilde{x}$ no son extremos opuestos de un espectro.

Debido a que los valores grandes presentes en la muestra del ejemplo 1.13 afectan a $\bar{x}$ más que $\tilde{x}$, $\tilde{x} < \bar{x}$ con esos datos. Aunque tanto $\bar{x}$ como $\tilde{x}$ ubican el centro de un conjunto de datos, en general no serán iguales porque se enfocan en aspectos diferentes de la muestra.

Análogo a $\tilde{x}$ como valor medio de la muestra es un valor medio de la población, la **mediana poblacional**, denotada por $\tilde{\mu}$. Como con $\bar{x}$ y μ, se puede pensar en utilizar la mediana muestral $\tilde{x}$ para hacer una inferencia sobre $\tilde{\mu}$. En el ejemplo 1.13, se podría utilizar $\tilde{x} = 10.05$ como estimación de la concentración de la mediana en toda la población de la cual se tomó la muestra. A menudo se utiliza una mediana para describir ingresos o salarios (debido a que no es influida en gran medida por unos pocos salarios grandes). Si el salario mediano de una muestra de ingenieros fuera $\tilde{x} = 66\,416$ dólares se podría utilizar como base para concluir que el salario mediano de todos los ingenieros es de más de 60 000 dólares.



a) Asimétrico negativo b) Simétrico c) Asimétrico positivo

Figura 1.15 Tres formas diferentes de una distribución de población.

La media μ y la mediana $\tilde{\mu}$ poblacionales en general no serán idénticas. Si la distribución de la población es positiva o negativamente asimétrica, como se ilustra en la figura 1.15, entonces $\mu \neq \tilde{\mu}$. Cuando éste es el caso, al hacer inferencias primero se debe decidir cuál de las dos características de la población es de mayor interés y luego proceder como corresponda.

Otras medidas de localización: cuartiles, percentiles y medias recortadas

La mediana (poblacional o muestral) divide el conjunto de datos en dos partes iguales. Para obtener medidas de ubicación más finas, se podrían dividir los datos en más de dos partes. Tentativamente, los cuartiles dividen el conjunto de datos en cuatro partes iguales y las observaciones arriba del tercer cuartil constituyen el cuarto superior del conjunto de datos, el segundo cuartil es idéntico a la mediana y el primer cuartil separa el cuarto inferior de los tres cuartos superiores. Asimismo, un conjunto de datos (muestra o población) puede ser incluso más finamente dividido por medio de percentiles, el 99º percentil separa el 1% más alto del 99% más bajo, y así sucesivamente. A menos que el número de observaciones sea un múltiplo de 100, se debe tener cuidado al obtener percentiles. En el capítulo 4 se utilizarán percentiles con conexión con ciertos modelos de poblaciones infinitas y por tanto su discusión se pospone hasta ese punto.

La media es bastante sensible a un solo valor extremo, mientras que la mediana es insensible a muchos valores apartados. Como el comportamiento extremo de uno u otro tipo podría ser indeseable, se consideran brevemente medidas alternativas que no son ni sensibles como $\bar{x}$ ni tan insensibles como $\tilde{x}$. Para motivar estas alternativas, obsérvese que $\bar{x}$ y $\tilde{x}$ se encuentran en extremos opuestos de la misma "familia" de medidas. La media es el promedio de todos los datos, mientras que la mediana resulta de eliminar todos excepto uno o dos valores medios y luego promediar. Parafraseando, la media implica recortar 0% de cada extremo de la muestra, mientras que en el caso de la mediana se recorta la cantidad máxima posible de cada extremo. Una **muestra recortada** es un término medio entre $\bar{x}$ y $\tilde{x}$. Una media 10% recortada, por ejemplo, se calcularía eliminando el 10% más pequeño y el 10% más grande de la muestra y luego promediando lo que queda.

Ejemplo 1.14 La producción de Bidri es una artesanía tradicional de India. Las artesanías Bidri (tazones, recipientes, etc.) se funden con una aleación que contiene principalmente zinc y algo de cobre. Considere las siguientes observaciones sobre contenido de cobre (%) de una muestra de artefactos Bidri tomada del Museo Victoria y Albert en Londres ("Enigmas of Bidri", *Surface Engr.*, 2005: 333-339), enlistadas en orden creciente.

2.0 2.4 2.5 2.6 2.6 2.7 2.7 2.8 3.0 3.1 3.2 3.3 3.3
3.4 3.4 3.6 3.6 3.6 3.6 3.7 4.4 4.6 4.7 4.8 5.3 10.1

La figura 1.16 es una gráfica de puntos de los datos. Una característica prominente es el valor extremo único en el extremo superior; la distribución está más dispersa en la región de valores grandes que en el caso de valores pequeños. La media muestral y la mediana son 3.65 y 3.35, respectivamente. Se obtiene una media recortada ($\bar{x}_r$) con un porcentaje de recorte de 100(2/26) = 7.7% al eliminar las dos observaciones más pequeñas y las dos más grandes; esto da

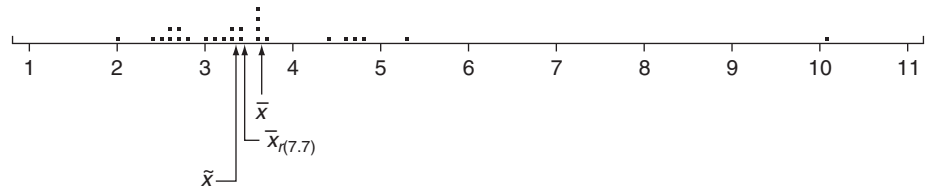


Figura 1.16 Gráfica de puntos de contenidos de cobre del ejemplo 1.14.

$\bar{x}_{r(7.7)} = 3.42$. El recorte en este caso elimina el valor extremo más grande y por tanto aproxima la media recortada hacia la mediana. ■

Una media recortada con un porcentaje de recorte moderado, algo entre 5 y 25%, producirá una medida del centro que no es ni tan sensible a los valores apartados como la media ni tan insensible como la mediana. Si el porcentaje de recorte deseado es $100\alpha\%$ y $n\alpha$ no es un entero, la media recortada debe ser calculada por interpolación. Por ejemplo, considérese $\alpha = 0.10$ para un porcentaje de recorte de 10% y $n = 26$ como en el ejemplo 1.14. Entonces $\bar{x}_{r(10)}$ sería el promedio ponderado apropiado de la media 7.7% recortada calculada allí y la media 11.5% recortada que resulta de recortar tres observaciones de cada extremo.

Datos categóricos y proporciones muestrales

Cuando los datos son categóricos, una distribución de frecuencia o una distribución de frecuencia relativa proporciona un resumen tabular efectivo de los datos. Las cantidades resumidas numéricas naturales en esta situación son las frecuencias individuales y las frecuencias relativas. Por ejemplo, si se realiza una encuesta de personas que poseen cámaras digitales para estudiar la preferencia de marcas y cada persona en la muestra identifica la marca de cámara que él o ella posee, con lo cual se podría contar el número que poseen Cannon, Sony, Kodak, y así sucesivamente. Considérese muestrear una población dividida en dos partes, una que consiste en sólo dos categorías (tal como votó o no votó en la última elección, si posee o no una cámara digital, etc.). Si x denota el número en la muestra que cae en la categoría 1, entonces el número en el categoría 2 es $n - x$. La frecuencia relativa o *proporción muestral* en la categoría 1 es x/n y la proporción muestral en la categoría 2 es $1 - x/n$. Que 1 denote una respuesta que cae en la categoría 1 y que 0 denote una respuesta que cae en la categoría 2. Un tamaño de muestra de $n = 10$ podría dar entonces las respuestas 1, 1, 0, 1, 1, 1, 0, 0, 1, 1. La media muestral de esta muestra numérica es (como el número de unos $= x = 7$)

$$\frac{x_1 + \cdots + x_n}{n} = \frac{1 + 1 + 0 + \cdots + 1 + 1}{10} = \frac{7}{10} = \frac{x}{n} = \text{proporción muestral}$$

Más generalmente, *enfóquese la atención en una categoría particular y codifíquense los resultados de modo que se anote un* 1 *para una observación comprendida en la categoría y un* 0 *para una observación no comprendida en la categoría. Entonces la proporción muestral de observaciones comprendida en la categoría es la media muestral de la secuencia de los* 1 *y los* 0. Por consiguiente se puede utilizar una media muestral para resumir los resultados de una muestra categórica. Estos comentarios también se aplican a situaciones en las cuales las categorías se definen agrupando valores en una muestra o población numérica (p. ej., podría existir interés en saber si las personas han tenido su automóvil actual durante por lo menos 5 años, en lugar de estudiar la duración exacta de la tenencia).

Análogo a la proporción muestral x/n de personas u objetos que caen en una categoría particular, que p represente la proporción de aquellos presentes en toda la población que cae en la categoría. Como con x/n, p es una cantidad entre 0 y 1 y mientras que x/n es una característica de muestra, p es una característica de la población. La relación entre las

dos es igual a la relación entre $\tilde{x}$ y $\tilde{\mu}$ y entre $\bar{x}$ y μ. En particular, subsecuentemente se utilizará x/n para hacer inferencias sobre p. Si, por ejemplo, una muestra de 100 propietarios de automóviles reveló que 22 tenían su automóvil desde por lo menos 5 años atrás, en tal caso se podría utilizar $22/100 = 0.22$ como estimación puntual de la proporción de todos los propietarios que tenían su automóvil desde por lo menos 5 años atrás. Se estudiarán las propiedades de x/n como una estimación de p para ver cómo se puede utilizar x/n para responder otras preguntas inferenciales. Con k categorías ($k > 2$), se pueden utilizar las k proporciones muestrales para responder preguntas sobre las proporciones de población $p_1, \ldots, p_k$.

EJERCICIOS Sección 1.3 (33-43)

33. El artículo ("The Pedaling Technique of Elite Endurance Cyclists", *Inst. J. of Sport Biomechanics*, 1991: 29-53) reportó los datos adjuntos sobre potencia de una sola pierna sometida a una alta carga de trabajo.

244 191 160 187 180 176 174
205 211 183 211 180 194 200

a. Calcule e interprete la media y la mediana muestral.
b. Suponga que la primera observación hubiera sido 204 en lugar de 244. ¿Cómo cambiarían la media y la mediana?
c. Calcule una media recortada eliminando las observaciones muestrales más pequeñas y más grandes. ¿Cuál es el porcentaje de recorte correspondiente?
d. El artículo también reportó valores de potencia de una sola pierna con carga de trabajo baja. La media muestral de $n = 13$ observaciones fue $\bar{x} = 119.8$ (en realidad 119.7692) y la 14a. observación, algo así como un valor extremo, fue 159. ¿Cuál es el valor de $\bar{x}$ de toda la muestra?

34. La exposición a productos microbianos, especialmente endotoxina, puede tener un impacto en la vulnerabilidad a enfermedades alérgicas. El artículo ("Dust Sampling Methods for Endotoxin-An Essential, But Underestimated Issue", *Indoor Air*, 2006: 20-27) consideró temas asociados con la determinación de concentración de endotoxina. Los siguientes datos sobre concentración (EU/mg) en polvo asentado de una muestra de hogares urbanos y otra de casas campestres fueron amablemente suministrados por los autores del artículo citado.

U: 6.0 5.0 11.0 33.0 4.0 5.0 80.0 18.0 35.0 17.0 23.0
C: 4.0 14.0 11.0 9.0 9.0 8.0 4.0 20.0 5.0 8.9 21.0 9.2 3.0 2.0 0.3

a. Determine la media muestral de cada muestra. ¿Cómo se comparan?
b. Determine la mediana muestral de cada muestra. ¿Cómo se comparan? ¿Por qué es la mediana de la muestra urbana tan diferente de la media de dicha muestra?
c. Calcule la media recortada de cada muestra eliminando la observación más pequeña y más grande. ¿Cuáles son los porcentajes de recorte correspondientes? ¿Cómo se comparan los valores de estas medias recortadas a las medias y medianas correspondientes?

35. La presión de inyección mínima (lb/pulg^2) de especímenes moldeados por inyección de fécula de maíz se determinó con ocho especímenes diferentes (la presión más alta corresponde a una mayor dificultad de procesamiento) y se obtuvieron las siguientes observaciones (tomadas de "Thermoplastic Starch Blends with Polyethylene-Co-Vinyl Alcohol: Processability and Physical Properties", *Polymer Engr. and Science*, 1994: 17-23):

15.0 13.0 18.0 14.5 12.0 11.0 8.9 8.0

a. Determine los valores de la media muestral, la mediana muestral y la media 12.5% recortada y compare estos valores.
b. ¿En cuánto se podría incrementar la observación de la muestra más pequeña, actualmente 8.0, sin afectar el valor de la mediana muestral?
c. Suponga que desea los valores de la media y la mediana muestrales cuando las observaciones están expresadas en kilogramos por pulgada cuadrada (kg/pulg^2) en lugar de lb/pulg^2. ¿Es necesario volver a expresar cada observación en kg/pulg^2 o se pueden utilizar los valores calculados en el inciso a) directamente? [*Sugerencia*: 1 kg = 2.2 lb.]

36. Una muestra de 26 trabajadores de plataforma petrolera marina tomaron parte en un ejercicio de escape y se obtuvieron los datos adjuntos de tiempo (s) para completar el escape ("Oxygen Consumption and Ventilation During Escape from an Offshore Platform", *Ergonomics*, 1997: 281-292):

389 356 359 363 375 424 325 394 402
373 373 370 364 366 364 325 339 393
392 369 374 359 356 403 334 397

a. Construya una gráfica de tallo y hojas de los datos. ¿Cómo sugiere la gráfica que la media y mediana muestrales se comparen?
b. Calcule los valores de la media y mediana muestrales [*Sugerencia*: $\sum x_i = 9638$.]
c. ¿En cuánto se podría incrementar el tiempo más largo, actualmente de 424, sin afectar el valor de la mediana muestral? ¿En cuánto se podría disminuir este valor sin afectar el valor de la mediana muestral?
d. ¿Cuáles son los valores de $\bar{x}$ y $\tilde{x}$ cuando las observaciones se reexpresan en minutos?

37. El artículo ("Snow Cover and Temperature Relationships in North America and Eurasia", *J. Climate and Applied Meteorology*, 1983: 460-469) utilizó técnicas estadísticas para relacionar la cantidad de cobertura de nieve sobre cada

continente para promediar la temperatura continental. Los datos allí presentados incluyeron las siguientes diez observaciones de la cobertura de nieve en octubre en Eurasia durante los años 1970-1979 (en millones de km^2):

6.5 12.0 14.9 10.0 10.7 7.9 21.9 12.5 14.5 9.2

¿Qué reportaría como valor representativo, o típico de cobertura de nieve en octubre durante este periodo y qué motivaría su elección?

38. Los valores de presión sanguínea a menudo se reportan a los 5 mmHg más cercanos (100, 105, 110, etc.). Suponga que los valores de presión sanguínea reales de nueve individuos seleccionados al azar son

118.6 127.4 138.4 130.0 113.7 122.0 108.3 131.5 133.2

a. ¿Cuál es la mediana de los valores de presión sanguínea *reportados*?

b. Suponga que la presión sanguínea del segundo individuo es 127.6 en lugar de 127.4 (un pequeño cambio en un solo valor). ¿Cómo afecta esto a la mediana de los valores reportados? ¿Qué dice esto sobre la sensibilidad de la mediana al redondeo o agrupamiento en los datos?

39. La propagación de grietas provocadas por fatiga en varias partes de un avión ha sido el tema de extensos estudios en años recientes. Los datos adjuntos se componen de vidas de propagación (horas de vuelo/10^4) para alcanzar un tamaño de agrietamiento dado en orificios para sujetadores utilizados en aviones militares ("Statistical Crack Propagation in Fastener Holes ander Spectrum Loading", *J. Aircraft*, 1983: 1028-1032):

0.736 0.863 0.865 0.913 0.915 0.937 0.983 1.007
1.011 1.064 1.109 1.132 1.140 1.153 1.253 1.394

a. Calcule y compare los valores de la media y mediana muestrales.

b. ¿En cuánto se podría disminuir la observación muestral más grande sin afectar el valor de la mediana?

40. Calcule la mediana muestral, media 25% recortada, media 10% recortada y media muestral de los datos de duración dados en el ejercicio 27 y compare estas medidas.

41. Se eligió una muestra de $n = 10$ automóviles y cada uno se sometió a una prueba de choque a 5 mph. Denotando un carro sin daños visibles por S (por éxito) y un carro con daños por F, los resultados fueron los siguientes:

S S F S S S F F S S

a. ¿Cuál es el valor de la proporción muestral de éxitos x/n?

b. Reemplace cada S con 1 y cada F con 0. Acto seguido calcule $\bar{x}$ de esta muestra numéricamente codificada. ¿Cómo se compara $\bar{x}$ con x/n?

c. Suponga que se decide incluir 15 carros más en el experimento. ¿Cuántos de éstos tendrían que ser S para dar $x/n = 0.80$ para toda la muestra de 25 carros?

42. **a.** Si se agrega una constante c a cada x_i en una muestra y se obtiene $y_i = x_i + c$, ¿cómo se relacionan la media y mediana muestrales de las y_i con la media y mediana muestrales de las x_i? Verifique sus conjeturas.

b. Si cada x_i se multiplica por una constante c y se obtiene $y_i = cx_i$, responda la pregunta del inciso a). De nuevo, verifique sus conjeturas.

43. Un experimento para estudiar la duración (en horas) de un cierto tipo de componente implicaba poner diez componentes en operación y observarlos durante 100 horas. Ocho de ellos fallaron durante dicho periodo y se registraron las duraciones. Denote las duraciones de dos componentes que continuaron funcionando después de 100 horas por 100+. Las observaciones muestrales resultantes fueron:

48 79 100+ 35 92 86 57 100+ 17 29

¿Cuáles de las medidas del centro discutidas en esta sección pueden ser calculadas y cuáles son los valores de dichas medidas? [*Nota*: Se dice que los datos obtenidos con este experimento están "censurados a la derecha".]

1.4 Medidas de variabilidad

El reporte de una medida de centro da sólo información parcial sobre un conjunto o distribución de datos. Diferentes muestras o poblaciones pueden tener medidas idénticas de centro y aún diferir entre sí en otras importantes maneras. La figura 1.17 muestra gráficas de puntos de tres muestras con las mismas media y mediana, aunque el grado de dispersión en

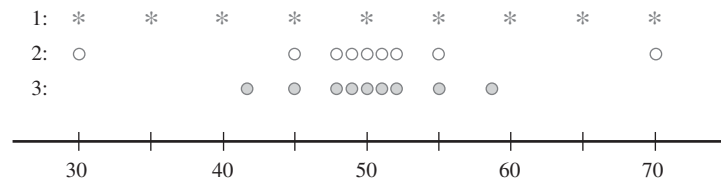


Figura 1.17 Muestras con medidas idénticas de centro pero diferentes cantidades de variabilidad.

torno al centro es diferente para las tres muestras. La primera tiene la cantidad más grande de variabilidad, la tercera tiene la cantidad más pequeña y la segunda es intermedia con respecto a las otras dos.

Medidas de variabilidad de datos muestrales

La medida más simple de variabilidad en una muestra es el **rango**, el cual es la diferencia entre los valores muestrales más grande y más pequeño. El valor del rango de la muestra 1 en la figura 1.17 es mucho más grande que el de la muestra 3, lo que refleja más variabilidad en la primera muestra que en la tercera. Un defecto del rango, no obstante, es que depende de sólo las dos observaciones más extremas y hace caso omiso de las posiciones de los $n - 2$ valores restantes. Las muestras 1 y 2 en la figura 1.17 tienen rangos idénticos, aunque cuando se toman en cuenta las observaciones entre los dos extremos, existe mucho menos variabilidad o dispersión en la segunda muestra que en la primera.

Las medidas principales de variabilidad implican las **desviaciones de la media**, $x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$. Es decir, las desviaciones de la media se obtienen restando $\bar{x}$ de cada una de la n observaciones muestrales. Una desviación será positiva si la observación es más grande que la media (a la derecha de la media sobre el eje de medición) y negativa si la observación es más pequeña que la media. Si todas las desviaciones son pequeñas en magnitud, entonces todas las x_i se aproximan a la media y hay poca variabilidad. Alternativamente, si algunas de las desviaciones son grandes en magnitud, entonces algunas x_i quedan lejos de $\bar{x}$ lo que sugiere una mayor cantidad de variabilidad. Una forma simple de combinar las desviaciones en una sola cantidad es promediarlas. Desafortunadamente, esto es una mala idea:

$$\text{suma de desviaciones} = \sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

por lo que la desviación promedio siempre es cero. La verificación utiliza varias reglas estándar y el hecho de que $\sum \bar{x} = \bar{x} + \bar{x} + \cdots + \bar{x} = n\bar{x}$:

$$\sum (x_i - \bar{x}) = \sum x_i - \sum \bar{x} = \sum x_i - n\bar{x} = \sum x_i - n\left(\frac{1}{n}\sum x_i\right) = 0$$

¿Cómo se puede evitar que las desviaciones negativas y positivas se neutralicen entre sí cuando se combinan? Una posibilidad es trabajar con los valores absolutos de las desviaciones y calcular la desviación absoluta promedio $\sum |x_i - \bar{x}|/n$. Como la operación de valores absolutos conduce a dificultades teóricas, considérense en cambio las desviaciones al cuadrado $(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, \ldots, (x_n - \bar{x})^2$. En vez de utilizar la desviación al cuadrado promedio $\sum (x_i - \bar{x})^2/n$, por varias razones se divide la suma de desviaciones al cuadrado entre $n - 1$ en lugar de entre n.

DEFINICIÓN

La **varianza muestral**, denotada por s^2 está dada por

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

La **desviación estándar muestral**, denotada por s, es la raíz cuadrada (positiva) de la varianza

$$s = \sqrt{s^2}$$

Obsérvese que s^2 y s son no negativas. La unidad de s es la misma que la de cada una de las x_i. Si por ejemplo, las observaciones son eficiencias de combustible en millas por galón, entonces se podría tener $s = 2.0$ mpg. Una interpretación preliminar de la desviación estándar

muestral es que es el tamaño de una desviación típica o representativa de la media muestral dentro de la muestra dada. Por tanto si $s = 2.0$ mpg, entonces algunas x_i en la muestra se aproximan más que 2.0 a $\bar{x}$, en tanto que otras están más alejadas; 2.0 es una desviación representativa (o "estándar") de la eficiencia de combustible media. Si $s = 3.0$ de una segunda muestra de carros de otro tipo, una desviación típica en esta muestra es aproximadamente 1.5 veces la de la primera muestra, una indicación de más variabilidad en la segunda muestra.

Ejemplo 1.15 La resistencia es una característica importante de los materiales utilizados en casas prefabricadas. Cada uno de $n = 11$ elementos de placa prefabricados se sometieron a prueba de esfuerzo severo y se registró el ancho máximo (mm) de las grietas resultantes. Los datos proporcionados (tabla 1.3) aparecieron en el artículo ("Prefabricated Ferrocement Ribbed Elements for Low-Cost Housing", *J. Ferrocement*, 1984: 347-364).

Tabla 1.3 Datos del ejemplo 1.15

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
0.684	−0.9841	0.9685
2.540	0.8719	0.7602
0.924	−0.7441	0.5537
3.130	1.4619	2.1372
1.038	−0.6301	0.3970
0.598	−1.0701	1.1451
0.483	−1.1851	1.4045
3.520	1.8519	3.4295
1.285	−0.3831	0.1468
2.650	0.9819	0.9641
1.497	−0.1711	0.0293
$\sum x_i = 18.349$	$\sum(x_i - \bar{x}) = -0.0001$	$S_{xx} = \sum(x_i - \bar{x})^2 = 11.9359$

$\bar{x} = 18.349/11 = 1.6681$

Los efectos de redondeo hacen que la suma de las desviaciones no sea exactamente cero. El numerador de s^2 es 11.9359, por consiguiente $s^2 = 11.9359/(11 - 1) = 11.9359/10 = 1.19359$ y $s = \sqrt{1.19359} = 1.0925$ mm. ■

Motivación para s^2

Para explicar el porqué del divisor $n - 1$ en s^2, obsérvese primero que en tanto que s^2 mide la variabilidad muestral, existe una medida de variabilidad en la población llamada *varianza poblacional*. Se utilizará σ^2 (el cuadrado de la letra griega sigma minúscula) para denotar la varianza poblacional y σ para denotar la desviación estándar poblacional (la raíz cuadrada de σ^2). Cuando la población es finita y se compone de N valores,

$$\sigma^2 = \sum_{i=1}^{N} (x_i - \mu)^2/N$$

la cual es el promedio de todas las desviaciones al cuadrado con respecto a la media poblacional (para la población, el divisor es N y no $N - 1$). En los capítulos 3 y 4 aparecen definiciones más generales de σ^2.

Así como $\bar{x}$ se utilizará para hacer inferencias sobre la media poblacional μ, se deberá definir la variancia muestral de modo que pueda ser utilizada para hacer inferencias sobre σ^2. Ahora obsérvese que σ^2 implica desviaciones cuadradas con respecto a la media poblacional μ. Si en realidad se conociera el valor de μ, entonces se podría definir la

varianza muestral como la desviación al cuadrado promedio de las x_i de la muestra con respecto a μ. Sin embargo, el valor de μ casi nunca es conocido, por lo que se debe utilizar el cuadrado de la suma de las desviaciones con respecto a $\bar{x}$. Pero *las x_i tienden a acercarse más a su valor promedio que el promedio poblacional μ, así que para compensar esto se utiliza el divisor $n - 1$ en lugar de n*. En otras palabras, si se utiliza un divisor n en la varianza muestral, entonces la cantidad resultante tendería a subestimar σ^2 (se producen valores demasiado pequeños en promedio), mientras que si se divide entre el divisor un poco más pequeño $n - 1$ se corrige esta subestimación.

Se acostumbra referirse a s^2 que está basada en $n - 1$ **grados de libertad** (gl o df, por sus siglas en inglés). Esta terminología se deriva del hecho de que aunque s^2 está basada en las n cantidades $x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$, éstas suman 0, por lo que al especificar los valores de cualquier $n - 1$ de las cantidades se determina el valor restante. Por ejemplo, si $n = 4$ y $x_1 - \bar{x} = 8$, $x_2 - \bar{x} = -6$ y $x_4 - \bar{x} = -4$, entonces automáticamente $x_3 - \bar{x} = 2$, así que sólo tres de los cuatro valores de $x_i - \bar{x}$ son libremente determinados (3 gl).

Una fórmula para calcular s^2

Es mejor obtener s^2 con software estadístico o bien utilizar una calculadora que permita ingresar datos en la memoria y luego ver s^2 con un solo golpe de tecla. Si su calculadora no tiene esta capacidad, existe una fórmula alternativa para S_{xx} que evita calcular las desviaciones. La fórmula implica sumar $(\sum x_i)^2$, sumar y luego elevar al cuadrado y $\sum x_i^2$, elevar al cuadrado y sumar.

Una alternativa para el numerador de s^2 es

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

Comprobación Como $\bar{x} = \sum x_i/n$, $n\bar{x}^2 = (\sum x_i)^2/n$. Entonces

$$\begin{aligned}\sum (x_i - \bar{x})^2 &= \sum (x_i^2 - 2\bar{x} \cdot x_i + \bar{x}^2) = \sum x_i^2 - 2\bar{x} \sum x_i + \sum (\bar{x})^2 \\ &= \sum x_i^2 - 2\bar{x} \cdot n\bar{x} + n(\bar{x})^2 = \sum x_i^2 - n(\bar{x})^2\end{aligned}$$

Ejemplo 1.16 La cantidad de luz reflejada por las hojas ha sido utilizada para varios propósitos, incluidas la evaluación del color del césped, la estimación del estado del nitrógeno y la medición de la biomasa. El artículo ("Leaf Reflectance-Nitrogen-Chlorophyll Relations in Buffel-Grass", *Photogrammetric Engr. and Remote Sensing*, 1985: 463-466) dio las siguientes observaciones obtenidas por medio de espectrofotogrametría, de la reflexión de las hojas en condiciones experimentales.

Observación	x_i	x_i^2	Observación	x_i	x_i^2
1	15.2	231.04	9	12.7	161.29
2	16.8	282.24	10	15.8	249.64
3	12.6	158.76	11	19.2	368.64
4	13.2	174.24	12	12.7	161.29
5	12.8	163.84	13	15.6	243.36
6	13.8	190.44	14	13.5	182.25
7	16.3	265.69	15	12.9	166.41
8	13.0	169.00			
				$\sum x_i = 216.1$	$\sum x_i^2 = 3168.13$

La fórmula de cálculo ahora da

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 3168.13 - \frac{(216.1)^2}{15}$$
$$= 3168.13 - 3113.28 = 54.85$$

con la cual $s^2 = S_{xx}/(n-1) = 54.85/14 = 3.92$ y $s = 1.98$. ■

Tanto la fórmula definitoria como la de cálculo para s^2 pueden ser sensibles al redondeo, por lo que en los cálculos intermedios se deberá usar tanta precisión decimal como sea posible.

Algunas otras propiedades de s^2 pueden mejorar el entendimiento y facilitar el cálculo.

PROPOSICIÓN

Sean $x_1, x_2, \ldots, x_n$ una muestra y c cualquier constante no cero.

1. Si $y_1 = x_1 + c, y_2 = x_2 + c, \ldots, y_n = x_n + c$, entonces $s_y^2 = s_x^2$, y

2. Si $y_1 = cx_1, \ldots, y_n = cx_n$, entonces $s_y^2 = c^2 s_x^2$, $s_y = |c| s_x$,

donde s_x^2 es la varianza muestral de las x y s_y^2 es la varianza muestral de las y.

En palabras, el resultado 1 dice que si se suma una constante c (o resta) de cada valor de dato, la varianza no cambia. Esto es intuitivo, puesto que la adición o sustracción de c cambia la localización del conjunto de datos pero deja las distancias iguales entre los valores de datos. De acuerdo con el resultado 2, la multiplicación de cada x_i por c hace que s^2 sea multiplicada por un factor de c^2. Estas propiedades pueden ser comprobadas al observar que en el resultado 1, $\bar{y} = \bar{x} + c$ y que en el resultado 2, $\bar{y} = c\bar{x}$.

Gráficas de caja

Las gráficas de tallo y hojas e histogramas transmiten impresiones un tanto generales sobre un conjunto de datos, mientras que un resumen único tal como la media o la desviación estándar se enfoca en sólo un aspecto de los datos. En años recientes, se ha utilizado con éxito un resumen gráfico llamado *gráfica de caja* para describir varias de las características más prominentes de un conjunto de datos. Estas características incluyen 1) el centro, 2) la dispersión, 3) el grado y naturaleza de cualquier alejamiento de la simetría y 4) la identificación de las observaciones "extremas o apartadas" inusualmente alejadas del cuerpo principal de los datos. Como incluso un solo valor extremo puede afectar drásticamente los valores de $\bar{x}$ y s, una gráfica de caja está basada en medidas "resistentes" a la presencia de unos cuantos valores apartados, la mediana y una medida de variabilidad llamada *dispersión de los cuartos*.

DEFINICIÓN

Se ordenan las observaciones de la más pequeña a la más grande y se separa la mitad más pequeña de la más grande; se incluye la mediana $\tilde{x}$ en ambas mitades si n es impar. En tal caso el **cuarto inferior** es la mediana de la mitad más pequeña y el **cuarto superior** es la mediana de la mitad más grande. Una medida de dispersión que es resistente a los valores apartados es la **dispersión de los cuartos** f_s, dada por

$$f_s = \text{cuarto superior} - \text{cuarto inferior}$$

En general, la dispersión de los cuartos no se ve afectada por las posiciones de las observaciones comprendidas en el 25% más pequeño o el 25% más grande de los datos. Por consiguiente es resistente a valores apartados.

La gráfica de caja más simple se basa en el siguiente resumen de cinco números:

x_i más pequeñas cuarto inferior mediana cuarto superior x_i más grandes

Primero, se traza una escala de medición horizontal. Luego se coloca un rectángulo sobre este eje; el lado izquierdo del rectángulo está en el cuarto inferior y el derecho en el cuarto superior (por lo que el ancho de la caja $= f_s$). Se coloca un segmento de línea vertical o algún otro símbolo dentro del rectángulo en la ubicación de la mediana; la posición del símbolo de mediana con respecto a los dos lados da información sobre asimetría en el 50% medio de los datos. Por último, se trazan "bigotes" hacia fuera de ambos extremos del rectángulo hacia las observaciones más pequeñas y más grandes. También se puede trazar una gráfica de caja con orientación vertical mediante modificaciones obvias en el proceso de construcción.

Ejemplo 1.17 Se utilizó ultrasonido para reunir los datos de corrosión adjuntos de la placa de piso de un tanque elevado utilizado para almacenar petróleo crudo ("Statistical Analysis of UT Corrosion Data from Floor Plates of a Crude Oil Aboveground Storage Tank", *Materials Eval.*, 1994: 846-849); cada observación es la profundidad de picadura más grande en la placa, expresada en milésimas de pulgada.

$$\underbrace{40\ 52\ 55\ 60\ 70\ 75\ 85\ 85\ 90}\ \overbrace{90\ 92\ 94\ 94\ 95\ 98\ 100\ 115\ 125\ 125}$$

El resumen de cinco números es como sigue:

x_i más pequeña $= 40$ cuarto inferior $= 72.5$ $\tilde{x} = 90$ cuarto superior $= 96.5$
x_i más grande $= 125$

La figura 1.18 muestra la gráfica de caja resultante. El lado derecho de la caja está mucho más cerca a la mediana que el izquierdo, lo que indica una asimetría sustancial en la mitad derecha de los datos. El ancho de la caja (f_s) también es razonablemente grande con respecto al rango de datos (distancia entre las puntas de los bigotes).

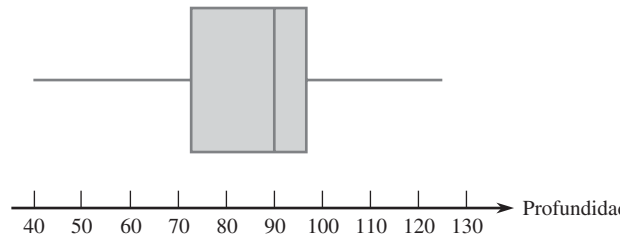


Figura 1.18 Gráfica de caja de los datos de corrosión.

La figura 1.19 muestra los resultados obtenidos con MINITAB en respuesta a la petición de describir los datos de corrosión. La media recortada es el promedio de las 17 observaciones que permanecen después de eliminar los valores más grandes y más pequeños (porcentaje de recorte $\approx 5\%$), Q1 y Q3 son los cuartiles inferior y superior; éstos son similares a los cuartos pero se calculan de una manera diferente; el error estándar promedio (SE Mean) es $s/\sqrt{n}$; esta será una importante cantidad en el trabajo subsiguiente con respecto a inferencias en torno a μ.

```
Profundidad  N  Media  Media  Media recortada  Desv. estándar  Media SE
   variable 19  86.32  90.00        86.76           23.32         5.35

Profundidad     Mínima     Máxima       Q1       Q3
   variable      40.00     125.00    70.00    98.00
```

Figura 1.19 Descripción de MINITAB de los datos de profundidad de picaduras. ■

Gráficas de caja que muestran valores apartados

Una gráfica de caja puede ser embellecida para indicar explícitamente la presencia de valores apartados. Muchos procedimientos inferenciales se basan en la suposición de que la distribución de la población es normal (un cierto tipo de curva en forma de campana). Incluso

DEFINICIÓN

Cualquier observación a más de $1.5f_s$ del cuarto más cercano es un **valor apartado (o atípico)**. Un valor apartado es **extremo** si se encuentra a más de $3f_s$ del cuarto más cercano y **moderado** de lo contrario.

un solo valor apartado extremo que aparezca en la muestra advierte al investigador que tales procedimientos pueden ser no confiables y la presencia de varios valores apartados transmite el mismo mensaje.

Modifíquese ahora la construcción previa de una gráfica de caja trazando un bigote que sale de cada extremo de la caja hacia las observaciones más pequeñas y más grandes que *no* son valores apartados. Cada valor apartado moderado está representado por un círculo cerrado y cada valor apartado extremo por uno abierto. Algunos programas de computadora estadísticos no distinguen entre valores apartados moderados y extremos.

Ejemplo 1.18 Los efectos de descargas parciales en la degradación de materiales para cavidades aislantes tienen implicaciones importantes en relación con las duraciones de componentes de alto voltaje. Considérese la siguiente muestra de $n = 25$ anchos de pulso de descargas lentas en una cavidad cilíndrica de polietileno. (Estos datos son consistentes con un histograma de 250 observaciones en el artículo "Assessment of Dielectric Degradation by Ultrawide-band PD Detection", *IEEE Trans. on Dielectrics and Elec. Insul.*, 1995: 744-760.) El autor del artículo señala el impacto de una amplia variedad de herramientas estadísticas en la interpretación de datos de descarga.

5.3	8.2	13.8	74.1	85.3	88.0	90.2	91.5	92.4	92.9	93.6	94.3	94.8
94.9	95.5	95.8	95.9	96.6	96.7	98.1	99.0	101.4	103.7	106.0	113.5	

Las cantidades pertinentes son

$$\tilde{x} = 94.8 \qquad \text{cuarto inferior} = 90.2 \qquad \text{cuarto superior} = 96.7$$
$$f_s = 6.5 \qquad 1.5f_s = 9.75 \qquad 3f_s = 19.50$$

Por lo tanto, cualquier observación menor que $90.2 - 9.75 = 80.45$ o mayor que $96.7 + 9.75 = 106.45$ es un valor apartado. Hay un valor apartado en el extremo superior de la muestra y cuatro en el extremo inferior. Debido a que $90.2 - 19.5 = 70.7$, las tres observaciones 5.3, 8.2 y 13.8 son valores apartados extremos; los otros dos son moderados. Los bigotes se extienden a 85.3 y 106.0, las observaciones más extremas que no son valores apartados. La gráfica de caja resultante aparece en la figura 1.20. Existe una gran cantidad de asimetría negativa en la mitad media de la muestra así como también en toda la muestra.

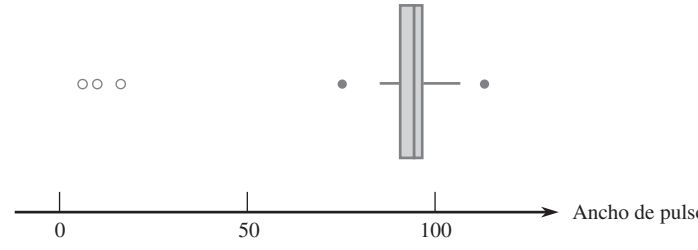


Figura 1.20 Gráfica de caja de los datos de ancho de pulso que muestra valores apartados moderados y extremos. ■

Gráficas de caja comparativas

Una gráfica de caja comparativa o lado a lado es una forma muy efectiva de revelar similitudes y diferencias entre dos o más conjuntos de datos compuestos de observaciones de la misma variable, observaciones de eficiencia de consumo de combustible de cuatro tipos distintos de automóviles, rendimientos de cosechas de tres variedades diferentes y así sucesivamente.

Ejemplo 1.19 En años recientes, algunas evidencias sugieren que las altas concentraciones de radón bajo techo pueden estar ligadas al desarrollo de cánceres en niños, pero muchos profesionales de la salud aún no están convencidos. Un artículo reciente ("Indoor Radon and Childhood Cancer", *The Lancet*, 1991: 1537-1538) presentó los datos adjuntos sobre concentración de radón (Bq/m^3) en dos muestras diferentes de casas. La primera consistió en casas en las cuales un niño diagnosticado con cáncer había estado residiendo. Las casas en la segunda muestra no incluían casos registrados de cáncer infantil. La figura 1.21 presenta una gráfica de tallo y hojas de los datos.

1. Con cáncer		2. Sin cáncer	
9683795	0	95768397678993	
86071815066815233150	1	12271713114	
12302731	2	99494191	
8349	3	839	
5	4		
7	5	55	
	6		
	7		Tallo: dígitos de decenas
HI: 210	8	5	Hojas: dígitos de unidades

Figura 1.21 Gráfica de tallo y hojas del ejemplo 1.19.

El resumen de cantidades numéricas es el siguiente:

	$\bar{x}$	$\tilde{x}$	s	f_s
Con cáncer	22.8	16.0	31.7	11.0
Sin cáncer	19.2	12.0	17.0	18.0

Los valores tanto de la media como de la mediana sugieren que la muestra de cáncer se encuentra en el centro un poco a la derecha de la muestra sin cáncer sobre la escala de medición. La media, sin embargo, exagera la magnitud de este desplazamiento, en gran medida debido a la observación 210 en la muestra con cáncer. Los valores de s sugieren más variabilidad en la muestra con cáncer que en la muestra sin cáncer, pero las dispersiones de los cuartos contradicen esta impresión. De nuevo, la observación 210, un valor apartado extremo, es el culpable. La figura 1.22 muestra una gráfica de caja comparativa generada por el

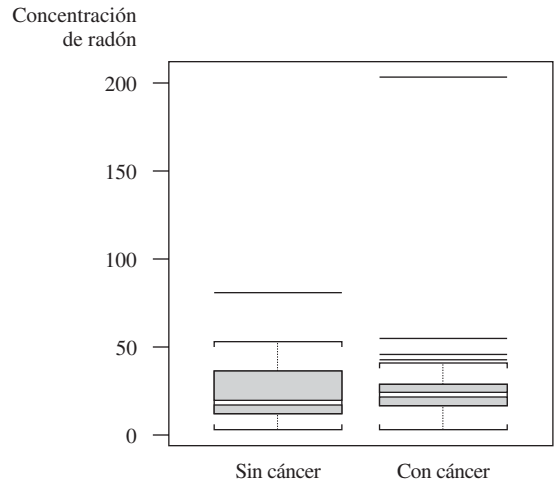


Figura 1.22 Gráfica de caja de los datos del ejemplo 1.19, obtenida con S-Plus.

programa de computadora S-Plus. La caja sin cáncer aparece alargada en comparación con la caja con cáncer ($f_s = 18$ *vs.* $f_s = 11$) y las posiciones de las líneas medianas en las dos cajas muestran más asimetría en la mitad media de la muestra sin cáncer que la muestra con cáncer. Los valores apartados están representados por segmentos de línea horizontales y no hay distinción entre los valores apartados moderados y extremos. ■

EJERCICIOS Sección 1.4 (44-61)

44. El artículo ("Oxygen Consumption During Fire Suppression: Error of Heart Rate Estimation", *Ergonomics*, 1991: 1469-1474) reportó los siguientes datos sobre consumo de oxígeno (ml/kg/min) para una muestra de diez bomberos que realizaron un simulacro de supresión de incendio.

29.5 49.3 30.6 28.2 28.0 26.3 33.9 29.4 23.5 31.6

Calcule lo siguiente:

a. El rango muestral.
b. La varianza muestral s^2 a partir de la definición (es decir, calculando primero las desviaciones y luego elevándolas al cuadrado, etcétera).
c. La desviación estándar muestral.
d. s^2 utilizando el método más corto.

45. Se determinó el valor del módulo de Young (GPa) de placas fundidas compuestas de ciertos sustratos intermetálicos y se obtuvieron las siguientes observaciones muestrales ("Strength and Modulus of a Molybdenum-Coated Ti-25A1-10Nb-3U-1Mo Intermetallic", *J. of Materials Engr. and Performance*, 1997: 46-50):

116.4 115.9 114.6 115.2 115.8

a. Calcule $\bar{x}$ y las desviaciones de la media.
b. Use las desviaciones calculadas en el inciso a) para obtener la varianza muestral y la desviación estándar muestral.
c. Calcule s^2 utilizando la fórmula para el numerador S_{xx}.
d. Reste 100 de cada observación para obtener una muestra de valores transformados. Ahora calcule la varianza muestral de estos valores transformados y compárela con s^2 de los datos originales.

46. Las observaciones adjuntas de viscosidad estabilizada (cP) realizadas en probetas de un cierto grado de asfalto con 18% de caucho agregado se tomaron del artículo ("Viscosity Characteristics of Rubber-Modified Asphalts", *J. of Materials in Civil Engr.* 1996: 153-156):

2781 2900 3013 2856 2888

a. ¿Cuáles son los valores de la media y mediana muestrales?
b. Calcule la varianza muestral por medio de la fórmula de cálculo. [*Sugerencia*: Primero reste un número conveniente de cada observación.]

47. Calcule e interprete los valores de la mediana muestral, la media muestral y la desviación estándar muestral de las siguientes observaciones de resistencia a la fractura (MPa, leídas en una gráfica que aparece en el artículo ("Heat-Resistant Active Brazing of Silicon Nitride: Mechanical Evaluation of Braze Joints", *Welding J.*, agosto de 1997):

87 93 96 98 105 114 128 131 142 168

48. El ejercicio 34 presentó los siguientes datos sobre concentración de endotoxina en polvo asentado, obtenidos con una muestra de casas urbanas y una muestra de casas campestres:

U: 6.0 5.0 11.0 33.0 4.0 5.0 80.0 18.0 35.0 17.0 23.0
C: 4.0 14.0 11.0 9.0 9.0 8.0 4.0 20.0 5.0 8.9 21.0 9.2 3.0 2.0 0.3

a. Determine el valor de la desviación estándar muestral de cada muestra, interprete estos valores y luego contraste la variabilidad en las dos muestras. [*Sugerencia*: $\sum x_i =$ 237.0 para la muestra urbana y = 128.4 para la muestra campestre y $\sum x_i^2 =$ 10 079 para la muestra urbana y 1617.94 para la muestra campestre.]
b. Calcule la dispersión de los cuartos de cada muestra y compare. ¿Transmiten el mismo mensaje las dispersiones de los cuartos sobre la variabilidad que las desviaciones estándar? Explique.
c. Los autores del artículo citado también proporcionan concentraciones de endotoxina en el polvo presente en bolsas captadoras de polvo:

U: 34.0 49.0 13.0 33.0 24.0 24.0 35.0 104.0 34.0 40.0 38.0 1.0
C: 2.0 64.0 6.0 17.0 35.0 11.0 17.0 13.0 5.0 27.0 23.0 28.0 10.0 13.0 0.2

Construya una gráfica de caja comparativa (como se hizo en el artículo citado) y compare y contraste las cuatro muestras.

49. Un estudio de la relación entre edad y varias funciones visuales (tales como agudeza y percepción de profundidad) reportó las siguientes observaciones de área de la lámina esclerótica (mm^2) de las cabezas del nervio óptico humano ("Morphometry of Nerve Fiber Bundle Pores in the Optic Nerve Head of the Human", *Experimental Eye Research*, 1988: 559-568):

2.75 2.62 2.74 3.85 2.34 2.74 3.93 4.21 3.88
4.33 3.46 4.52 2.43 3.65 2.78 3.56 3.01

a. Calcule $\sum x_i$ y $\sum x_i^2$.
b. Use los valores calculados en el inciso a) para calcular la varianza muestral s^2 y luego la desviación estándar muestral s.

50. En 1997, una mujer demandó a un fabricante de teclados de computadora y lo acusó de que sus repetitivas lesiones por esfuerzo eran provocadas por el teclado (*Genessy v. Digital*

Equipment Corp.). El jurado adjudicó $3.5 millones por el dolor y sufrimiento pero la corte anuló dicha adjudicación por considerarla una compensación irrazonable. Al hacer esta determinación, la corte identificó un grupo "normativo" de 27 casos similares y especificó una adjudicación razonable como una dentro de dos desviaciones estándar de la media de las adjudicaciones en los 27 casos. Las 27 adjudicaciones fueron (en el rango de los $1000) 37, 60, 75, 115, 135, 140, 149, 150, 238, 290, 340, 410, 600, 750, 750, 750, 1050, 1100, 1139, 1150, 1200, 1200, 1250, 1576, 1700, 1825 y 2000 con las cuales $\sum x_i = 20179$, $\sum x_i^2 = 24\,657\,511$. ¿Cuál es la cantidad máxima posible que podría ser adjudicada conforme a la regla de dos desviaciones estándar?

51. El artículo ("A Thin-Film Oxygen Uptake Test for the Evaluation of Automotive Crankcase Lubricants", *Lubric. Engr.*, 1984: 75-83) reportó los siguientes datos sobre tiempo de inducción de oxidación (min) de varios aceites comerciales:

87 103 130 160 180 195 132 145 211 105 145
153 152 138 87 99 93 119 129

a. Calcule la varianza muestral y la desviación estándar.
b. Si las observaciones se volvieran a expresar en horas, ¿cuáles serían los valores resultantes de la varianza de la muestra y la desviación estándar muestral? Responda sin realizar en realidad la reexpresión.

52. Las primeras cuatro desviaciones de la media en una muestra de $n = 5$ tiempos de reacción fueron 0.3, 0.9, 1.0 y 1.3. ¿Cuál es la quinta desviación de la media? Dé una muestra para la cual estas son las cinco desviaciones de la media.

53. Reconsidere los datos sobre el área de lámina esclerótica dados en el ejercicio 49.
a. Determine los cuartos inferior y superior.
b. Calcule el valor de la dispersión de los cuartos.
c. Si los dos valores muestrales más grandes, 4.33 y 4.52 hubieran sido 5.33 y 5.52, ¿cómo afectaría esto a f_s? Explique.
d. ¿En cuánto se podría incrementar la observación 2.34 sin afectar a f_s? Explique.
e. Si la 18a. observación, $x_{18} = 4.60$, se suma a la muestra, ¿cuál es f_s?

54. Considere las siguientes observaciones sobre resistencia al esfuerzo cortante (MPa) de una junta unida de una manera particular (tomadas de una gráfica que aparece en el artículo ("Diffusion of Silicon Nitride to Austenitic Stainless Steel without Interlayers", *Metallurgical Trans*., 1993: 1835-1843).

22.2 40.4 16.4 73.7 36.6 109.9
30.0 4.4 33.1 66.7 81.5

a. ¿Cuáles son los valores de los cuartos y cuál es el valor de f_s?
b. Construya una gráfica de caja basada en el resumen de cinco números y comente sobre sus características.
c. ¿Qué tan grande o pequeña tiene que ser una observación para calificar como valor apartado? ¿Como valor apartado extremo?
d. ¿En cuánto podría disminuir la observación más grande sin afectar f_s?

55. He aquí una gráfica de tallo y hojas de los datos de tiempo de escape introducidos en el ejercicio 36 de este capítulo.

```
32 | 55
33 | 49
34 |
35 | 6699
36 | 34469
37 | 03345
38 | 9
39 | 2347
40 | 23
41 |
42 | 4
```

a. Determine el valor de la dispersión de los cuartos.
b. ¿Hay algunos valores apartados en la muestra? ¿Algunos valores apartados extremos?
c. Construya una gráfica de caja y comente sobre sus características.
d. ¿En cuánto se podría disminuir la observación más grande, actualmente de 424, sin afectar el valor de la dispersión de los cuartos?

56. Se determinó la cantidad de contaminación por aluminio (ppm) en plástico de cierto tipo con una muestra de 26 probetas de plástico y se obtuvieron los siguientes datos ("The Lognormal Distribution for Modeling Quality Data when the Mean Is Near Zero", *J. of Quality Technology*, 1990: 105-110):

30 30 60 63 70 79 87 90 101
102 115 118 119 119 120 125 140 145
172 182 183 191 222 244 291 511

Construya una gráfica de caja que muestre valores apartados y comente sobre sus características.

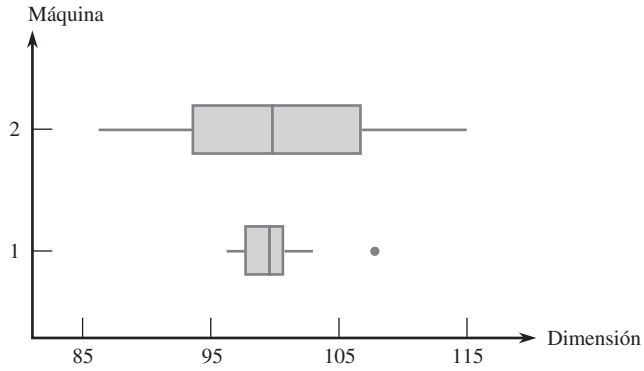
57. Se seleccionó una muestra de 20 botellas de vidrio de un tipo particular y se determinó la resistencia a la presión interna de cada botella. Considere la siguiente información parcial sobre la muestra:
mediana = 202.2 cuarto inferior = 196.0
cuarto superior = 216.8
Las tres observaciones más pequeñas 125.8 188.1 193.7
Las tres observaciones más grandes 221.3 230.5 250.2

a. ¿Hay valores apartados en la muestra? ¿Algunos valores apartados extremos?
b. Construya una gráfica de caja que muestre valores apartados y comente sobre cualesquiera características interesantes.

58. Una compañía utiliza dos máquinas diferentes para fabricar piezas de cierto tipo. Durante un solo turno, se obtuvo una muestra de $n = 20$ piezas producidas por cada máquina y se determinó el valor de una dimensión crítica particular de cada pieza. La gráfica de caja comparativa que aparece en la parte superior de la página 41 se construyó con los datos resultantes. Compare y contraste las dos muestras.

59. Se determinó la concentración de cocaína (mg/l) tanto con una muestra de individuos que murieron de delirio excitado (DE) inducido por el consumo de cocaína y con una muestra de aquellos que murieron de una sobredosis de cocaína sin delirio excitado; el tiempo de sobrevivencia de las personas

Gráfica de caja comparativa del ejercicio 58



en ambos grupos fue a lo sumo de 6 horas. Los datos adjuntos se tomaron de una gráfica de caja comparativa incluida en el artículo ("Fatal Excited Delirium Following Cocaine Use", *J. of Forensic Sciences*, 1997: 25-31).

Con DE 0 0 0 0 0.1 0.1 0.1 0.1 0.2 0.2 0.3 0.3
0.3 0.4 0.5 0.7 0.8 1.0 1.5 2.7 2.8
3.5 4.0 8.9 9.2 11.7 21.0

Sin DE 0 0 0 0 0 0.1 0.1 0.1 0.1 0.2 0.2 0.2
0.3 0.3 0.3 0.4 0.5 0.5 0.6 0.8 0.9 1.0
1.2 1.4 1.5 1.7 2.0 3.2 3.5 4.1
4.3 4.8 5.0 5.6 5.9 6.0 6.4 7.9
8.3 8.7 9.1 9.6 9.9 11.0 11.5
12.2 12.7 14.0 16.6 17.8

a. Determine las medianas, cuartos y dispersiones de los cuartos de las dos muestras,

b. ¿Existen algunos valores apartados en una u otra muestra? ¿Algunos valores apartados extremos?

c. Construya una gráfica de caja comparativa y utilícela como base para comparar y contrastar las muestras con DE y sin DE.

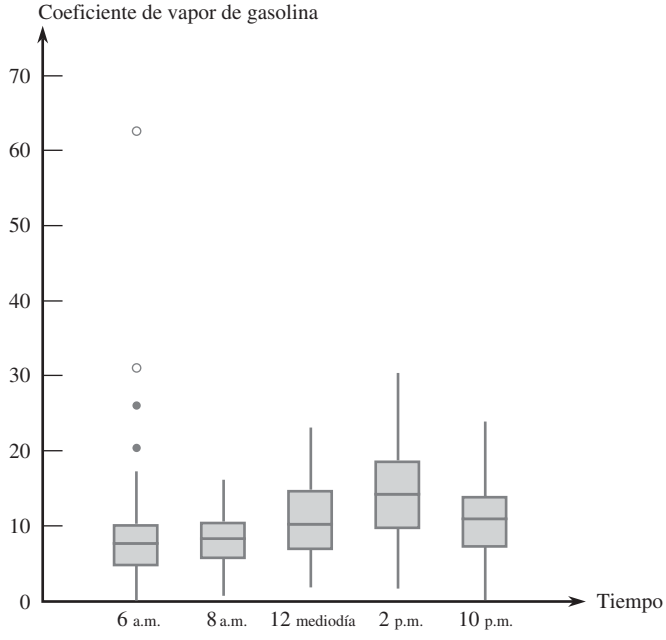
60. Se obtuvieron observaciones de resistencia al estallamiento (lb/pulg^2) tanto con soldaduras de cierre de toberas de prueba como con soldaduras para toberas de envases de producción ("Proper Procedures Are the Key to Welding Radioactive Waste Cannisters", *Welding J.*, agosto de 1997: 61-67).

Prueba	7200	6100	7300	7300	8000	7400
	7300	7300	8000	6700	8300	
Envase	5250	5625	5900	5900	5700	6050
	5800	6000	5875	6100	5850	6600

Construya una gráfica de caja comparativa y comente sobre las características interesantes (el artículo citado no incluía tal gráfica, pero los autores comentaron que habían visto uno.)

61. La gráfica de caja comparativa adjunta de coeficientes de vapor de gasolina de vehículos en Detroit apareció en el artículo ("Receptor Modeling Approach to VOC Emission Inventory Validation", *J. of Envir. Engr.*, 1995: 483-490). Discuta las características interesantes.

Gráfica de caja comparativa del ejercicio 61



EJERCICIOS SUPLEMENTARIOS (62-83)

62. Considere la siguiente información sobre resistencia a la tensión final (lb/pulg) de una muestra de $n = 4$ probetas de alambre de cobre al zirconio duro (de "Characterization Methods for Fine Copper Wire", *Wire J. Intl*., agosto de 1997: 74-80):

$\bar{x} = 76\,831$ $s = 180$, x_i más pequeña $= 76\,683$,
x_i más grande $= 77\,048$.

Determine los valores de las dos observaciones muestrales intermedias (¡pero no lo haga mediante conjeturas sucesivas!)

63. La cantidad de radiación recibida en un invernadero desempeña un importante papel al determinar el coeficiente de fotosíntesis. Las observaciones adjuntas sobre radiación solar incidente se leyeron en una gráfica que aparece en el artículo ("Radiation Components over Bare Planted Soils in a Greenhouse", *Solar Energy*, 1990: 1011-1016).

6.3	6.4	7.7	8.4	8.5	8.8	8.9
9.0	9.1	10.0	10.1	10.2	10.6	10.6
10.7	10.7	10.8	10.9	11.1	11.2	11.2
11.4	11.9	11.9	12.2	13.1		

Use algunos de los métodos estudiados en este capítulo para describir y resumir estos datos.

64. Los siguientes datos sobre emisiones de HC y CO de un vehículo particular se dieron en la introducción del capítulo.

*HC (g/*milla*)*	13.8	18.3	32.2	32.5
*CO (g/*milla*)*	118	149	232	236

a. Calcule las desviaciones estándar muestrales de las observaciones de HC y CO. ¿Parece justificarse la creencia difundida?

b. El *coeficiente de variación muestral* $s/\bar{x}$ (o $100\,s/\bar{x}$) evalúa el grado de variabilidad con respecto a la media. Los valores de este coeficiente para varios conjuntos de datos diferentes pueden ser comparados para determinar cuáles conjuntos de datos exhiben más o menos variación. Realice la comparación con los datos dados.

65. La distribución de frecuencia adjunta de observaciones de resistencia a la fractura (MPa) de barras de cerámicas cocidas en un horno particular apareció en el artículo ("Evaluating Tunnel Kiln Performance", *Amer. Ceramic Soc. Bull.*, agosto de 1997: 59-63).

Frecuencia de clase	81–<83	83–<85	85–<87	87–<89	89–<91
	6	7	17	30	43
Frecuencia de clase	91–<93	93–<95	95–<97	97–<99	
	28	22	13	3	

a. Construya un histograma basado en frecuencias relativas y comente sobre cualesquiera características interesantes.

b. ¿Qué proporción de las observaciones de resistencia son por lo menos de 85? ¿Menores que 95?

c. Aproximadamente, ¿qué proporción de las observaciones son menores que 90?

66. Una deficiencia de indicios de selenio en la dieta puede impactar negativamente el crecimiento, la inmunidad, la función muscular y neuromuscular y la fertilidad. La introducción de suplementos de selenio en vacas lecheras se justifica cuando las pasturas contienen niveles bajos de selenio. Los autores del artículo ("Effects of Short-Term Supplementation with Selenised Yeast on Milk Production and Composition of Lactating Cows", *Australian J. of Dairy Tech*., 2004: 199-203) suministraron los siguientes datos sobre la concentración de selenio en la leche (mg/l) obtenidos con una muestra de vacas a las que se les administró un suplemento de selenio y una muestra de control de vacas a las que no se les administró suplemento, tanto inicialmente como después de un periodo de 9 días.

Obs.	Se inicial	Cont. inicial	Se final	Cont. final
1	11.4	9.1	138.3	9.3
2	9.6	8.7	104.0	8.8
3	10.1	9.7	96.4	8.8
4	8.5	10.8	89.0	10.1
5	10.3	10.9	88.0	9.6
6	10.6	10.6	103.8	8.6
7	11.8	10.1	147.3	10.4
8	9.8	12.3	97.1	12.4
9	10.9	8.8	172.6	9.3
10	10.3	10.4	146.3	9.5
11	10.2	10.9	99.0	8.4
12	11.4	10.4	122.3	8.7
13	9.2	11.6	103.0	12.5
14	10.6	10.9	117.8	9.1
15	10.8		121.5	
16	8.2		93.0	

a. ¿Parecen ser similares las concentraciones iniciales de Se en las muestras de suplemento y en las de control? Use varias técnicas de este capítulo para resumir los datos y responder la pregunta planteada.

b. De nuevo use métodos de este capítulo para resumir los datos y luego describa cómo los valores de concentración de Se finales en el grupo de tratamiento difieren de aquellos en el grupo de control.

67. *Estenosis aórtica* se refiere al estrechamiento de la válvula aórtica en el corazón. El artículo ("Correlation Analysis of Stenotic Aortic Valve Flow Patterns Using Phase Constrast MRI", *Annals of Biomed. Engr*., 2005: 878-887) dio los siguientes datos sobre el diámetro de la raíz aórtica (cm) y el género de una muestra de pacientes con varios grados de estenosis aórtica:

H: 3.7 3.4 3.7 4.0 3.9 3.8 3.4 3.6 3.1 4.0 3.4 3.8 3.5
M: 3.8 2.6 3.2 3.0 4.3 3.5 3.1 3.1 3.2 3.0

a. Compare y contraste los diámetros observados en los dos géneros.

b. Calcule una media 10% recortada de cada una de las dos muestras y compare las demás medidas centrales (de la muestra de hombre, se debe utilizar el método de interpolación mencionado en la sección 1.3).

68. **a.** ¿Con qué valor de c es mínima la cantidad $\sum(x_i - c)^2$? [*Sugerencia*: Tome la derivada con respecto a c, iguale a 0 y resuelva.]

b. Utilizando el resultado del inciso a), ¿cuál de las dos cantidades $\sum(x_i - \bar{x})^2$ y $\sum(x_i - \mu)^2$ será más pequeña que la otra (suponiendo que $\bar{x} \neq \mu$)?

69. **a.** Sean a y b constantes y sea $y_i = ax_i + b$ con $i = 1, 2, \ldots, n$. ¿Cuáles son las relaciones entre $\bar{x}$ y $\bar{y}$ y entre s_x^2 y s_y^2?

b. Una muestra de temperaturas para iniciar una cierta reacción química dio un promedio muestral (°C) de 87.3 y una desviación estándar muestral de 1.04. ¿Cuáles son el promedio muestral y la desviación estándar medidos en °F? [*Sugerencia*: $F = \frac{9}{5}\ C + 32$.]

70. El elevado consumo de energía durante el ejercicio continúa después de que termina la sesión de entrenamiento. Debido a que las calorías quemadas por ejercicio contribuyen a la pérdida de peso y tienen otras consecuencias, es importante entender el proceso. El artículo ("Effect of Weight Training Exercise and Treadmill Exercise on Post-Exercise Oxygen Consumption", *Medicine and Science in Sports and Exercise*, 1998: 518-522) reportó los datos adjuntos tomados de un estudio en el cual se midió el consumo de oxígeno (litros) de forma continua durante 30 minutos de cada uno de 15 sujetos tanto después de un entrenamiento con pesas como después de una sesión de ejercicio en una caminadora.

Sujeto	1	2	3	4	5	6	7	8	9
	10	11	12	13	14	15			
Peso (x)	14.6	14.4	19.5	24.3	16.3	22.1			
	23.0	18.7	19.0	17.0	19.1	19.6			
	23.2	18.5	15.9						
Caminadora (y)	11.3	5.3	9.1	15.2	10.1	19.6			
	20.8	10.3	10.3	2.6	16.6	22.4			
	23.6	12.6	4.4						

a. Construya una gráfica de caja comparativa de las observaciones del ejercicio con pesas y la caminadora y comente sobre lo que ve.

b. Debido a que estos datos aparecen en pares (x, y), con mediciones de x y y de la misma variable en dos condiciones distintas, es natural enfocarse en las diferencias que existen en ellos: $d_1 = x_1 - y_1, \ldots, d_n = x_n - y_n$. Construya una gráfica de caja de las diferencias muestrales. ¿Qué sugiere la gráfica?

71. La siguiente es una descripción dada por MINITAB de los datos de resistencia dados en el ejercicio 13.

```
                                          Med.  Desv. Media
Resistencia  N   Media Mediana rec.  est.   SE
variable    153 135.39 135.40 135.41 4.59  0.37

Resistencia Mínima     Máxima        Q1        Q3
variable    122.20     147.70    132.95    138.25
```

a. Comente sobre cualesquiera características interesantes (los cuartiles y los cuartos son virtualmente idénticos en este caso).

b. Construya una gráfica de caja de los datos basada en los cuartiles y comente sobre lo que ve.

72. Los desórdenes y síntomas de ansiedad con frecuencia pueden ser tratados exitosamente con benzodiazepina. Se sabe que los animales expuestos a estrés exhiben una disminución de la ligadura de receptor de benzodiazepina en la corteza frontal. El artículo ("Decreased Benzodiazepine Receptor Binding in Prefrontal Cortex in Combat-Related Posttraumatic Stress Disorder", *Amer. J. of Psychiatry*. 2000: 1120-1126) describió el primer estudio de ligadura de receptor de benzodiazepina en individuos que sufren de PTSD. Los datos anexos sobre una medición de ligadura a receptor (volumen de distribución ajustado) se leyeron en una gráfica que aparece en el artículo.

PTSD: 10, 20, 25, 28, 31, 35, 37, 38, 38, 39, 39, 42, 46

Saludables: 23, 39, 40, 41, 43, 47, 51, 58, 63, 66, 67, 69, 72

Use varios métodos de este capítulo para describir y resumir los datos.

73. El artículo ("Can We Really Walk Straight?, *Amer. J. of Physical Anthropology*, 1992: 19-27) reportó sobre un experimento en el cual a cada uno de 20 hombres saludables se les pidió que caminarán en línea recta como fuera posible hacia un punto a 60 m de distancia a velocidad normal. Considérense las siguientes observaciones de cadencia (número de pasos por segundo):

0.95	0.85	0.92	.95	0.93	0.86	1.00	0.92	0.85	0.81
0.78	0.93	0.93	1.05	0.93	1.06	1.06	0.96	0.81	0.96

Use los métodos desarrollados en este capítulo para resumir los datos; incluya una interpretación o discusión en los casos en que sea apropiado. [*Nota*: El autor del artículo utilizó un análisis estadístico un tanto complejo para concluir que las personas no pueden caminar en línea recta y sugirió varias explicaciones para esto.]

74. La **moda** de un conjunto de datos numéricos es el valor que ocurre con más frecuencia en el conjunto.

a. Determine la moda de los datos de cadencia dados en el ejercicio 73.

b. Para una muestra categórica, ¿cómo definiría la categoría modal?

75. Se seleccionaron especímenes de tres tipos diferentes de cable y se determinó el límite de fatiga (Mpa) de cada espécimen y se obtuvieron los datos adjuntos.

Tipo 1	350	350	350	358	370	370	370	371
	371	372	372	384	391	391	392	
Tipo 2	350	354	359	363	365	368	369	371
	373	374	376	380	383	388	392	
Tipo 3	350	361	362	364	364	365	366	371
	377	377	377	379	380	380	392	

a. Construya una gráfica de caja comparativa y comente sobre las similitudes y diferencias.

b. Construya un diagrama de caja comparativo (una gráfica de puntos de cada muestra con una escala común). Comente sobre las similitudes y diferencias.

c. ¿Da la gráfica de caja comparativa del inciso a) una evaluación informativa de similitudes y diferencias? Explique su razonamiento.

76. Las tres medidas de centro introducidas en este capítulo son las media, la mediana y la media recortada. Dos medidas de centro adicionales que de vez en cuando se utilizan son el *rango medio*, el cual es el promedio de las observaciones más pequeñas y más grandes y el *cuarto medio*, el cual es el promedio de los dos cuartos. ¿Cuál de estas medidas de centro son resistentes a los efectos de los valores apartados y cuáles no? Explique su razonamiento.

77. Considere los siguientes datos sobre el tiempo de reparación activo (horas) de una muestra de $n = 46$ receptores de comunicaciones aerotransportados:

0.2	0.3	0.5	0.5	0.5	0.6	0.6	0.7	0.7	0.7	0.8	0.8
0.8	1.0	1.0	1.0	1.0	1.1	1.3	1.5	1.5	1.5	1.5	2.0
2.0	2.2	2.5	2.7	3.0	3.0	3.3	3.3	4.0	4.0	4.5	4.7
5.0	5.4	5.4	7.0	7.5	8.8	9.0	10.3	22.0	24.5		

Construya lo siguiente:

a. Una gráfica de tallo y hojas en la cual los dos valores más grandes se muestran por separado en la fila HI.

b. Un histograma basado en seis intervalos de clase con 0 como el límite inferior del primer intervalo y anchos de intervalo de 2, 2, 2, 4, 10 y 10, respectivamente.

78. Considere una muestra $x_1, x_2, \ldots, x_n$ y suponga que los valores de $\bar{x}$, s^2 y s han sido calculados.

a. Sea $y_i = x_i - \bar{x}$ con $i = 1, \ldots, n$. ¿Cómo se comparan los valores de s^2 y s de las y_i con los valores correspondientes de las x_i? Explique.

b. Sea $z_i = (x_i - \bar{x})/s$ con $i = 1, \ldots, n$. ¿Cuáles son los valores de la varianza muestral y la desviación estándar muestral de las z_i?

79. Si $\bar{x}_n$ y s_n^2 denotan la media y la varianza de la muestra $x_1, \ldots, x_n$ y si $\bar{x}_{n+1}$ y s_{n+1}^2 denotan estas cantidades cuando se agrega una observación adicional x_{n+1} a la muestra.

a. Demuestre cómo se puede calcular $\bar{x}_{n+1}$ con $\bar{x}_n$ y x_{n+1}.

b. Demuestre que

$$ns_{n+1}^2 = (n-1)s_n^2 + \frac{n}{n+1}(x_{n+1} - \bar{x}_n)^2$$

de modo que s_{n+1}^2 pueda ser calculada con x_{n+1}, $\bar{x}_n$, y s_n^2.

c. Suponga que una muestra de 15 torzales de hilo para telas dio por resultado un alargamiento del hilo mediano muestral de 12.58 mm y una desviación estándar muestral de 0.512 mm. ¿Cuáles son los valores de la media muestral y la desviación estándar muestral de las 16 observaciones de alargamiento?

80. Las distancias de recorrido de rutas de autobuses de cualquier sistema de tránsito particular por lo general varían de una ruta a otra. El artículo ("Planning of City Bus Routes", *J. of the Institution of Engineers*, 1995: 211-215) da la siguiente información sobre las distancias (km) de un sistema particular.

Distancia	6–<8	8–<10	10–<12	12–<14	14–<16
Frecuencia	6	23	30	35	32
Distancia	16–<18	18–<20	20–<22	22–<24	24–<26
Frecuencia	48	42	40	28	27
Distancia	26–<28	28–<30	30–<35	35–<40	40–<45
Frecuencia	26	14	27	11	2

a. Trace un histograma correspondiente a estas frecuencias.

b. ¿Qué proporción de estas distancias de ruta son menores que 20? ¿Qué proporción de estas rutas tienen distancias de recorrido de por lo menos 30?

c. ¿Aproximadamente cuál es el valor de 90º percentil de la distribución de distancia de recorrido de las rutas?

d. ¿Aproximadamente cuál es la distancia de recorrido de ruta mediana?

81. Un estudio realizado para investigar la distribución de tiempo de frenado total (tiempo de reacción más tiempo de movimiento de acelerador a freno, en ms) durante condiciones de manejo reales a 60 km/h da la siguiente información sobre la distribución de los tiempos ("A Field Study on Braking Response during Driving", *Ergonomics*, 1995: 1903-1910):

media = 535 mediana = 500 moda = 500
Desv. estd. = 96 mínima = 220 máxima = 925
5º percentil = 400 10º percentil = 430
90º percentil = 640 95º percentil = 720

¿Qué puede concluir sobre la forma de un histograma de estos datos? Explique su razonamiento.

82. Los datos muestrales $x_1, x_2, \ldots, x_n$ en ocasiones representan una **serie de tiempo**, donde x_t = el valor observado de una variable de respuesta x en el tiempo t. A menudo la serie observada muestra una gran cantidad de variación aleatoria, lo que dificulta estudiar el comportamiento a largo plazo. En tales situaciones, es deseable producir una versión alisada de la serie. Una técnica para hacerlo implica el **alisamiento o atenuación exponencial**. Se elige el valor de una constante de alisamiento α $(0 < \alpha < 1)$. Luego con $\bar{x}_t$ = valor alisado o atenuado en el tiempo t se hace $\bar{x}_1 = x_1$ con $t = 2, 3, \ldots, n$, $\bar{x}_t = \alpha x_t + (1 - \alpha)\bar{x}_{t-1}$.

a. Considere la siguiente serie de tiempo en la cual x_t = temperatura (°F) del efluente en una planta de tratamiento de aguas negras en el día t: 47, 54, 53, 50, 46, 46, 47, 50, 51, 50, 46, 52, 50, 50. Trace cada x_t contra t en un sistema de coordenadas de dos dimensiones (una gráfica de tiempo-serie). ¿Parece haber algún patrón?

b. Calcule las $\bar{x}_t$ con $\alpha = 0.1$. Repita con $\alpha = 0.5$. ¿Qué valor de α da una serie $\bar{x}_t$ más atenuada?

c. Sustituya $\bar{x}_{t-1} = \alpha x_{t-1} + (1 - \alpha)\bar{x}_{t-2}$ en el miembro de la derecha de la expresión para $\bar{x}_t$, acto seguido sustituya $\bar{x}_{t-2}$ en función de x_{t-2}, y $\bar{x}_{t-3}$, y así sucesivamente. ¿De cuántos de los valores $x_1, x_{t-1}, \ldots, x_1$ depende $\bar{x}_t$? ¿Qué le sucede al coeficiente de x_{t-k} conforme k se incrementa?

d. Remítase al inciso c). Si t es grande, ¿qué tan sensible es $\bar{x}_t$ a la inicialización $\bar{x}_1 = x_1$? Explique.

[*Nota*: Una referencia pertinente es el artículo "Simple Statistics for Interpreting Environmental Data", *Water Pollution Control Fed. J.*, 1981: 167-175.]

83. Considere las observaciones numéricas $x_1, \ldots, x_n$. Con frecuencia interesa saber si las x_i están (por lo menos en forma aproximada) simétricamente distribuidas en torno al mismo valor. Si n es por lo menos grande de manera moderada, el grado de simetría puede ser valorado con una gráfica de tallo y hojas o un histograma. Sin embargo, si n no es muy grande, las gráficas mencionadas no son informativas en

particular. Considere la siguiente alternativa. Que y_1 denote la x_i más pequeña, y_2 la segunda x_i más pequeña y así sucesivamente. Luego coloque los siguientes pares como puntos en una sistema de coordenadas de dos dimensiones $(y_n - \tilde{x}, \tilde{x} - y_1)$, $(y_{n-1} - \tilde{x}, \tilde{x} - y_2)$, $(y_{n-2} - \tilde{x}, \tilde{x} - y_3), \ldots$ Existen $n/2$ puntos cuando n es par y $(n-1)/2$ cuando n es impar.

a. ¿Qué apariencia tiene esta gráfica cuando la simetría en los datos es perfecta? ¿Qué apariencia tiene cuando las observaciones se alargan más sobre la mediana que debajo de ella (una larga cola superior)?

b. Los datos adjuntos sobre cantidad de lluvia (acres-pies) producida por 26 nubes bombardeadas se tomaron del artículo ("A Bayesian Analysis of Multiplicative Treatment Effect in Weather Modification", *Technometrics*, 1975: 161-166). Construya la gráfica y comente sobre el grado de simetría o la naturaleza del alejamiento de la misma.

4.1	7.7	17.5	31.4	32.7	40.6	92.4
115.3	118.3	119.0	129.6	198.6	200.7	242.5
255.0	274.7	274.7	302.8	334.1	430.0	489.1
703.4	978.0	1656.0	1697.8	2745.6		

Bibliografía

Chambers, John, William Cleveland, Beat Kleiner y Paul Tukey, *Graphical Methods for Data Analysis*, Brooks/Cole, Pacific Grove, CA, 1983. Una presentación altamente recomendada de varias metodologías gráficas y pictóricas en estadística.

Cleveland, William, *Visualizing Data*, Hobart Press, Summit, NJ, 1993. Un entretenido recorrido de técnicas pictóricas.

Devore, Jay y Roxy Peck, *Statistics*: *The Exploration and Analysis of Data* (5a. ed.), Thomson Brooks/Cole, Belmont, CA, 2005. Los primeros capítulos hacen un recuento no muy matemático de métodos para describir y resumir datos.

Freedman, David, Robert Pisani y Roger Purves, *Statistics* (3a. ed.), Norton, Nueva York, 1998. Un excelente estudio no muy matemático de razonamiento y metodología estadísticos básicos.

Hoaglin, David, Frederick Mosteller y John Tukey, *Understanding Robust and Exploratory Data Analysis*, Wiley, Nueva York, 1983. Discute el porqué y cómo deben ser utilizados los métodos exploratorios; es bueno por lo que se refiere a los detalles de gráficas de tallo y hojas y gráficas de caja.

Moore, David y William Notz, *Statistics: Concepts and Controversies* (6a. ed.), Freeman, San Francisco, 2006. Un libro de pasta blanda extremadamente fácil de leer y ameno que contiene una discusión intuitiva de problemas conectados con experimentos de muestreo y diseñados.

Peck, Roxy y colaboradores (eds.), *Statistics: A Guide to the Unknown* (4a. ed.), Thomson Brooks/Cole, Belmont, CA, 2006. Contiene muchos artículos no técnicos que describen varias aplicaciones de estadística.

Verzani, John, *Using R for Introductory Statistics*, Chapman y Hall/CRC, Boca Ratón, FL, 2005. Una introducción muy agradable al paquete de "software" R.