La figura 1.6 muestra una gráfica de puntos para los datos. Los puntos correspondientes a algunos valores muy cercanos (por ejemplo, 28.6 y 28.7) se han apilado verticalmente para evitar la aglomeración. Hay claramente una enorme variabilidad de un estado a otro. El valor más alto, para D.C., es obviamente un extremo atípico, y los otros cuatro valores en el extremo superior de los datos son candidatos a valores atípicos leves (MA, MN, Nueva York y ND). También hay un grupo de estados en el extremo inferior, situado principalmente en el sur y el suroeste. El porcentaje global para todo el país es de 39.3%; este no es un promedio simple de los 51 números, sino un promedio ponderado por tamaño de la población.

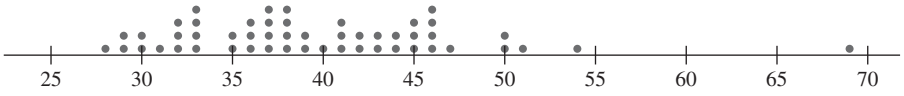


Figura 1.6 Gráfica de puntos para los datos del ejemplo 1.8

Una gráfica de puntos puede ser bastante enfadosa de construir y se ve muy saturada cuando el número de observaciones es grande. La siguiente técnica es muy adecuada en estas situaciones.

Histogramas

Para determinar el valor de una variable algunos datos numéricos se obtienen contando (el número de citatorios de tráfico que una persona recibió durante el año pasado, el número de personas que solicitan empleo durante un periodo específico), mientras que otros datos se obtienen tomando mediciones (el peso de un individuo, el 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 si se puede enumerar en una secuencia infinita (una en la cual exista un primer número, un segundo número y así sucesivamente). Una variable numérica es **continua** si sus valores posibles abarcan un intervalo completo sobre la recta numérica.

Una variable discreta x casi siempre resulta de haber contado, en cuyo caso los 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, 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 puedan determinar el pH, el tiempo de reacción, la altura y la concentración con un número arbitrariamente grande de decimales. Sin embargo, con la perspectiva de crear modelos matemáticos de distribuciones de datos, conviene imaginar todo un conjunto continuo de valores posibles.

Considere los datos compuestos de las observaciones de una variable discreta x. La **frecuencia** de cualquier valor particular x 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 las veces que ocurre el valor:

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

Suponga, 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 son 3, entonces

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

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

Si se multiplica una frecuencia relativa por 100 se obtiene un porcentaje; en el ejemplo de los 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 debido al redondeo. Una **distribución de frecuencia** es una tabla con las frecuencias o las frecuencias relativas, o ambas.

Construcción de un histograma para datos discretos

En primer lugar, se determinan 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: Los rectángulos deben medir lo mismo de ancho.

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

EJEMPLO 1.9 ¿Qué tan inusual es un juego de béisbol sin *hit* o de un solo *hit* en las ligas mayores y con qué frecuencia 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 y por cada uno de los juegos de nueve episodios que se jugaron entre 1989 y 1993.

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

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 mayores) que hacia la izquierda, un ligero "asimétrico positivo".

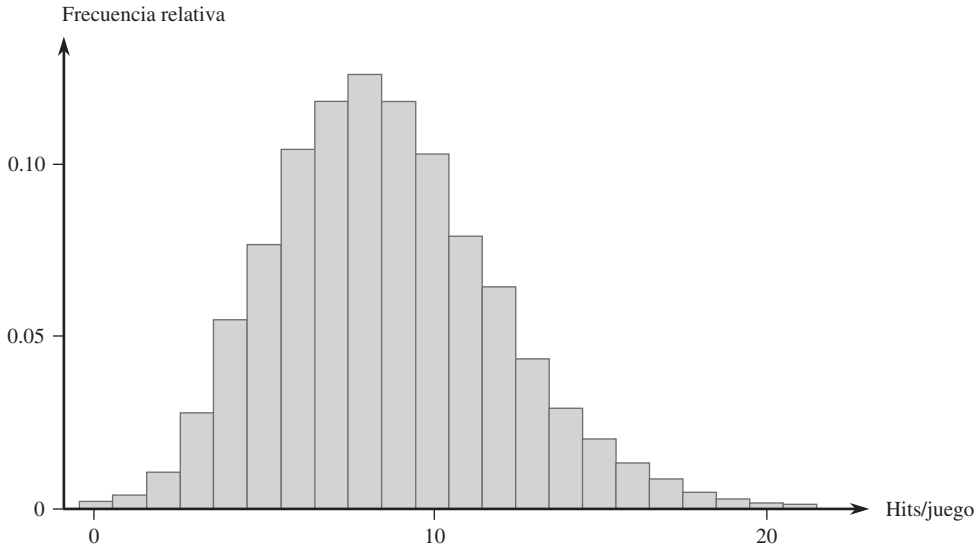


Figura 1.7 Histograma del número de hits por juego de nueve entradas

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

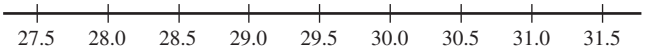
$$\begin{aligned}\text{proporción de juegos de dos hits a lo sumo} &= \begin{matrix}\text{frecuencia}\\ \text{relativa}\\ \text{para } x=0\end{matrix} + \begin{matrix}\text{frecuencia}\\ \text{relativa}\\ \text{para } x=1\end{matrix} + \begin{matrix}\text{frecuencia}\\ \text{relativa}\\ \text{para }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

De manera similar,

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

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

La construcción de un histograma para datos continuos (mediciones) implica subdividir el eje de medición entre un número adecuado de **intervalos de clase** o **clases**, de tal suerte que cada observación quede contenida exactamente en una clase. Suponga, por ejemplo, que se hacen 50 observaciones de x = eficiencia de consumo de combustible de un automóvil (mpg), la menor de las cuales es 27.8 y la mayor 31.4. Se podrían utilizar los límites de clase 27.5, 28.0, 28.5, … y 31.55 como se muestra a continuación:



Una dificultad potencial es que de vez en cuando una observación está en un límite de clase, por consiguiente, no cae exactamente en un intervalo, por ejemplo, 29.0. Una forma de tratar 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: clases con ancho igual

Se determinan 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 1.10 Las compañías generadoras de electricidad requieren información sobre el consumo de los clientes para obtener pronósticos precisos de la demanda. Investigadores de Wisconsin Power and Light determinaron el consumo de energía (en BTU) durante un periodo particular con una muestra de 90 hogares que utilizan gas. Se calculó un valor de consumo ajustado como sigue:

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

Esto dio como resultado los siguientes datos (una parte del conjunto de datos guardados FURNACE.MTW está disponible en Minitab), los cuales se ordenaron 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

En la figura 1.8 la característica del histograma que más llama la atención es su parecido a una curva en forma de campana, con el punto de simetría aproximadamente en 10.

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

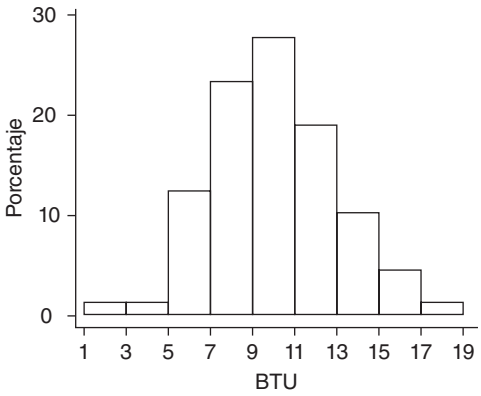


Figura 1.8 Histograma de los datos de consumo de energía del ejemplo 1.10

De acuerdo con el histograma,

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

La frecuencia relativa para la clase 9–<11 es aproximadamente 0.27, entonces se estima que aproximadamente la mitad de esta, o 0.135, queda entre 9 y 10. Por tanto,

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

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

No existen reglas inviolables en cuanto al número de clases o a la selección de las mismas. Entre 5 y 20 será satisfactorio 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 utilizarse. Una regla empírica razonable es

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

Es posible que las clases con ancho igual no sean una opción sensible si hay regiones en la escala de medición con una alta concentración de valores y otras donde los datos son muy escasos. La figura 1.9 muestra una gráfica de puntos de dicho conjunto de datos; hay una alta concentración en el medio y relativamente pocas observaciones que se extienden a ambos lados. Con un pequeño número de clases con ancho igual, casi todas las observaciones quedan exactamente en una o dos de las clases. Si se utiliza un número grande de clases con ancho igual, las frecuencias de muchas clases serán cero. Una buena opción es utilizar intervalos más anchos cerca de las observaciones extremas e intervalos más angostos en la región de alta concentración.

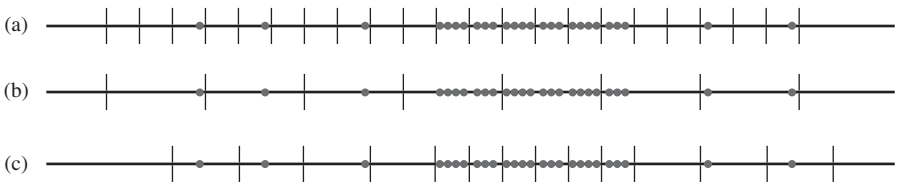


Figura 1.9 Selección de intervalos de clase para datos de "densidad variable": (a) intervalos de ancho igual muy cortos; (b) algunos intervalos de ancho igual; (c) intervalos de ancho desigual

Construcción de un histograma para datos continuos: clases con ancho desigual

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

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

Las alturas del rectángulo resultante se conocen usualmente como *densidades* y la escala vertical es la **escala de densidades**. Esta prescripción también funcionará cuando las clases tengan anchos iguales.

EJEMPLO 1.11 La corrosión del acero de refuerzo es un serio problema en las estructuras de concreto en ambientes afectados por condiciones climáticas severas. Por ello los investigadores han analizado el uso de barras de refuerzo fabricadas de un material compuesto. Se realizó un estudio para desarrollar directrices para adherir barras de refuerzo reforzadas con fibra de vidrio al concreto (**"Design Recommendations for Bond of GFRP Rebars to Concrete",** ***J. of Structural Engr*****., 1996: 247-254**). Considere las siguientes 48 observaciones de mediciones de fuerza adhesiva:

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

Clase	2 –<4	4–<6	6–<8	8–<12	12–<20	20–<30
Frecuencia	9	15	5	9	8	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 se muestra 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.

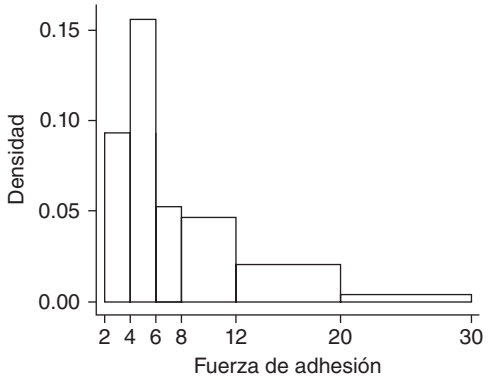


Figura 1.10 Histograma Minitab de densidad para la fuerza de adhesión del ejemplo 1.11 ■

Cuando las clases tienen anchos desiguales, sin utilizar una escala de densidades se obtendrá una gráfica con áreas distorsionadas. Para clases con anchos iguales el divisor es el mismo en cada cálculo de densidad y la aritmética adicional simplemente implica cambiar la escala en 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 la densidad por el ancho de clase, se obtiene

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

Es decir, *el área de cada rectángulo es la frecuencia relativa de la clase correspondiente*. Además, puesto que 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 papel importante 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, considere un gran conjunto de datos compuesto de los tiempos de manejo de automóviles en el trayecto entre San Luis Obispo, California y Monterey, California (sin contar el tiempo que se utilice para visitar lugares de interés, en comer, etc.). Este histograma mostraría dos crestas, una para los autos que toman la ruta interior (aproximadamente 2.5 horas) y otra para los que recorren la costa (3.5-4 horas). La bimodalidad no se presenta automáticamente en dichas situaciones. Sólo si los dos distintos histogramas están "muy alejados" respecto a sus dispersiones, la bimodalidad ocurrirá en el histograma de datos combinados. Por consiguiente, un conjunto de datos grande compuesto de las estaturas de los estudiantes universitarios no producirá un histograma bimodal porque la altura típica de los hombres, que aproximadamente es de 69 pulgadas, no está demasiado por encima de la altura típica de las mujeres, que es aproximadamente de 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, más probable es que se manifiesten bimodalidad o multimodalidad.

EJEMPLO 1.12 La figura 1.11(a) muestra un histograma Minitab de los pesos (en libras, lb) de los 124 jugadores que figuraban en las listas de los 49's de San Francisco y de los Patriotas de Nueva Inglaterra (equipos que al autor le gustaría ver reunidos en el Súper Tazón) el 20 de noviembre de 2009. La figura 1.11(b) es un histograma suavizado (que en realidad se llama *densidad estimada*) de los datos del paquete de software R. Tanto el histograma como el histograma suavizado muestran tres picos diferentes; el primero a la derecha es para los *linieros*, el del centro corresponde al peso de los *apoyadores* y el pico de la izquierda es para todos los demás jugadores (receptores abiertos, mariscales de campo, etc.).

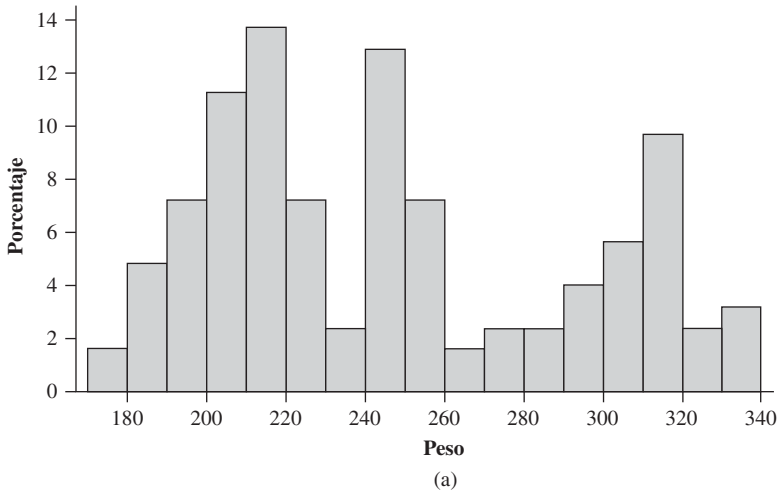


Figura 1.11 Peso de los jugadores de la NFL. (a) histograma y (b) histograma suavizado

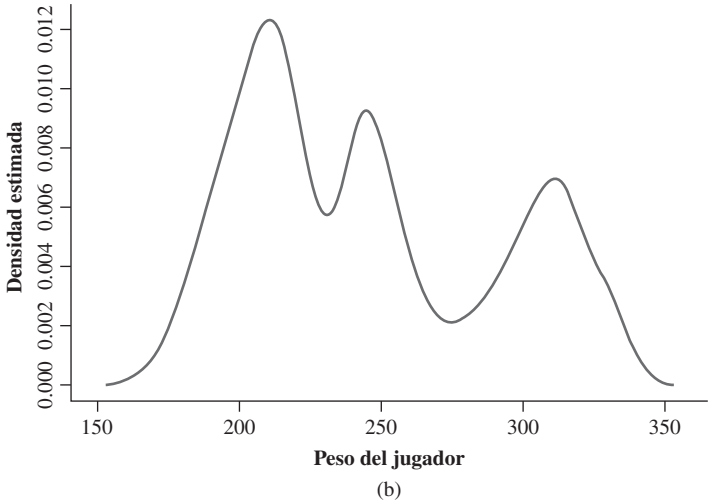


Figura 1.11 (continuación)

Un histograma es **simétrico** si la mitad izquierda es una imagen en espejo de la mitad derecha. Un histograma unimodal 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 1.12 muestra histogramas "suavizados", que se obtuvieron superponiendo una curva suavizada sobre los rectángulos e ilustran las varias posibilidades.

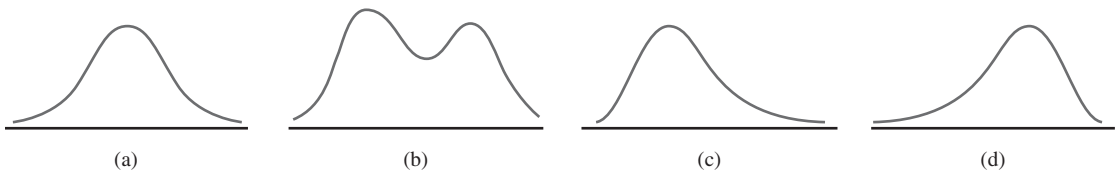


Figura 1.12 Histogramas suavizados: (a) unimodal simétrico; (b) bimodal; (c) positivamente asimétrico y (d) negativamente asimétrico

Datos cualitativos

Tanto una distribución de frecuencia como 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, de segundo, de tercero, de cuarto y graduados, mientras que en otros casos el orden será arbitrario, por ejemplo, católico, judío, protestante, etcétera. Con estos datos categóricos los intervalos sobre los cuales se construyen los rectángulos deberán ser de ancho igual.

EJEMPLO 1.13 El **Public Policy Institute of California** realizó una encuesta telefónica entre 2501 residentes adultos durante abril de 2006 para indagar lo que 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 1.13 muestra el histograma correspondiente (gráfica de barras).

A B D A A F C A C B E B A C
F D B C D A A C B E B C E A
B A A A B C C D F D B B A F
C B A C B E E D A B C E A A
F C B D D D B D C A F A A B
D E A E D B C A F A C D D A
A B A F D C A C B F D A E A
C D

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 tal modo que en el extremo izquierdo aparezca la categoría con la frecuencia más grande, enseguida 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 acumulada** y la frecuencia relativa acumulada de un intervalo de clase particular son la suma de las frecuencias y las frecuencias relativas, respectivamente, del intervalo y todos los intervalos que quedan debajo de él. Si, por ejemplo, tenemos cuatro intervalos con frecuencias 9, 16, 13 y 12, entonces las frecuencias acumuladas serán 9, 25, 38 y 50; y las frecuencias relativas acumuladas serán 0.18, 0.50, 0.76 y 1.00. Calcule las frecuencias acumuladas y las frecuencias relativas acumuladas de los datos del ejercicio 24.

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

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

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

b. ¿Qué proporción de cargas de fuego es menor de 600? ¿Y al menos menor de 1200?

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

1.3 Medidas de ubicación

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 y la interpretación de medidas resumidas numéricas. Es decir, se trata de extraer varios números resumidos a partir de los datos, 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 respecto a los datos categóricos.

Suponga, 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 ubicación y en particular su centro. Esta sección presenta métodos para describir la ubicación 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 el promedio aritmético del conjunto. Como casi siempre pensaremos que los números x_i constituyen una muestra, a menudo se hará referencia al promedio aritmético como la *media muestral* y se la denotará mediante $\bar{x}$.

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, si 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 1.14 En los últimos años ha habido un creciente interés comercial en el uso de lo que se conoce como *concreto internamente curado*. Este concreto comúnmente tiene inclusiones porosas en forma de agregado ligero (LWA). El artículo **"Characterizing Lightweight Aggregate Desorption at High Relative Humidities Using a Pressure Plate Apparatus" (*J. of Materials in Civil Engr*, 2012: 961-969)** reporta sobre un estudio en el cual los investigadores examinaron diversas propiedades físicas de 14 especímenes LWA. Estos son los porcentajes de absorción de agua de los especímenes durante 24 horas:

$x_1 = 16.0$	$x_2 = 30.5$	$x_3 = 17.7$	$x_4 = 17.5$	$x_5 = 14.1$
$x_6 = 10.0$	$x_7 = 15.6$	$x_8 = 15.0$	$x_9 = 19.1$	$x_{10} = 17.9$
$x_{11} = 18.9$	$x_{12} = 18.5$	$x_{13} = 12.2$	$x_{14} = 6.0$	

La figura 1.14 muestra una gráfica de puntos de los datos; un porcentaje de absorción de agua en medio de la decena entre diez y veinte parece ser "típico". Con $\sum x_i = 229.0$, la media muestral es

$$\bar{x} = \frac{229.0}{14} = 16.36$$

Una interpretación física de la media muestral nos indica cómo se evalúa el centro de una muestra. Cada punto en la gráfica de puntos se considera la representación de un peso de 1 lb. Entonces un punto de apoyo colocado con su punta en el eje horizontal estará en equilibrio precisamente cuando se encuentra en $\bar{x}$ (véase la figura 1.14). Por lo que la media muestral puede considerarse el punto de equilibrio de la distribución de las observaciones.

Figura 1.14 Gráfica de puntos de los datos del ejemplo 1.14 ■

Así como $\bar{x}$ representa el valor promedio de las observaciones incluidas en una muestra, es posible calcular el promedio de todos los valores de la población. Este promedio se conoce como la **media de la población** y se denota por la letra griega μ. Cuando existen N valores de la población (una población finita), entonces μ = (suma de los valores de población N)/N. En los capítulos 3 y 4 se dará una definición más general de μ, que se aplica a poblaciones tanto finitas como (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. Una de nuestras primeras tareas en inferencia estadística será presentar métodos basados en la media muestral para sacar conclusiones respecto una media de población. Por ejemplo, podríamos usar la media muestral $\bar{x} = 16.36$ calculada en el ejemplo 1.14 como una *estimación puntual* (un solo número que es nuestra "mejor" conjetura) de μ = el porcentaje de absorción de agua promedio verdadera para todos los especímenes tratados como se describe.

La media sufre de una deficiencia que, en algunas circunstancias, la convierte en una medida inapropiada del centro: su valor puede ser afectado en gran medida por la presencia incluso de un solo valor extremo (una observación inusualmente grande o pequeña). Por ejemplo, si en una muestra hay nueve empleados que ganan \$50 000 al año y un empleado cuyo salario anual es de \$150 000, el salario promedio de la muestra es \$60 000; en realidad este valor no parece representar los datos. En estas situaciones es conveniente recurrir a una medida menos sensible a los valores de $\bar{x}$ y por el momento propondremos una. Sin embargo, aunque $\bar{x}$ sí tiene este defecto potencial sigue siendo la medida más ampliamente utilizada, básicamente porque existen muchas poblaciones para las cuales un valor atípico 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 media 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 $x_1, x_2, \ldots, x_n$, 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,

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

EJEMPLO 1.15 Quienes no están familiarizados con la música clásica pueden creer que las instrucciones de un compositor para la reproducción de una pieza en particular son tan específicas que la duración no depende en absoluto del los intérpretes. Sin embargo, normalmente hay mucho espacio para la interpretación y para que los directores de orquesta y músicos puedan sacar el máximo provecho de ello. El autor se dirigió al sitio web **ArkivMusic.com** y seleccionó una muestra de 12 grabaciones de la Sinfonía # 9 de Beethoven ("Coral", una obra impresionante y hermosa), y generó las duraciones siguientes (en minutos) clasificadas en orden creciente:

62.3 62.8 63.6 65.2 65.7 66.4 67.4 68.4 68.8 70.8 75.7 79.0

He aquí una gráfica de puntos de los datos:

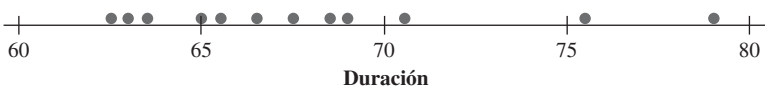


Figura 1.15 Gráfica de puntos de los datos para el ejemplo 1.15

Puesto que $n = 12$ es par, la mediana de la muestra es el promedio de los $n/2 = 6^{o}$ y $(n/2 + 1) = 7^{o}$ valores de la lista ordenada:

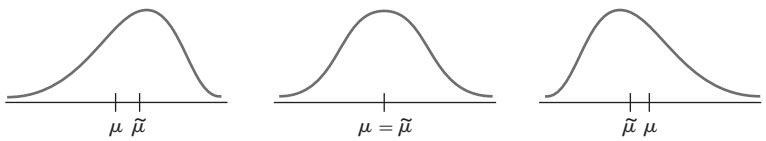
$$\tilde{x} = \frac{66.4 + 67.4}{2} = 66.90$$

Note que si la observación más grande, 79.0, no hubiera aparecido en la muestra, la mediana muestral resultante de las $n = 11$ observaciones restantes habría sido el valor medio 66.4 (el $[n + 1]/2 = 6^{o}$ valor ordenado, es decir, el sexto valor contado desde cualquier extremo de la lista ordenada). La media muestral es $\bar{x} = \sum x_i = 816.1/12 = 68.01$, la cual es poco más de un minuto más grande que la mediana. La media se sale un poco respecto a la mediana, ya que la muestra "se extiende" un poco más en el extremo superior que en el extremo inferior. ■

Los datos del ejemplo 1.15 ilustran una importante propiedad de $\tilde{x}$ en contraste con $\bar{x}$. La mediana muestral es muy insensible a los valores atípicos. Si, por ejemplo, las dos x_i más grandes se incrementan desde 75.7 y 79.0 hasta 85.7 y 89.0, respectivamente, $\tilde{x}$ no se vería afectada. Por tanto, en el tratamiento de valores atípicos, $\bar{x}$ y $\tilde{x}$ no son extremos opuestos de un espectro. Ambas cantidades describen el lugar donde se centran los datos, pero en general no serán iguales porque se enfocan en aspectos diferentes de la muestra.

Análogo a $\tilde{x}$ como valor medio de la muestra existe un valor medio de la población, la **mediana poblacional**, denotada por $\tilde{\mu}$. Tal como con $\bar{x}$ y μ, puede pensarse en utilizar la mediana muestral $\tilde{x}$ para hacer una inferencia sobre $\tilde{\mu}$. En el ejemplo 1.15 se podría utilizar $\tilde{x} = 66.90$ como una estimación de la mediana de tiempo para la población de todas las grabaciones.

La media μ y la mediana $\tilde{\mu}$ poblacionales en general no serán idénticas. Si la distribución de la población es positivamente o negativamente asimétrica, como se ilustra en la figura 1.16, entonces $\mu \neq \tilde{\mu}$. Cuando es este 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.



(a) Sesgo negativo (b) Simétrica (c) Sesgo positivo

Figura 1.16 Tres formas diferentes de distribución de la población

Otras medidas de ubicación: 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 dividen 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 mediante percentiles, el 99º percentil separa el más alto 1% del más bajo 99% 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 en conexión con ciertos modelos de poblaciones infinitas.

La media es bastante sensible a un solo valor extremo mientras que la mediana es insensible a muchos valores atípicos. Puesto que 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 observe 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 **media recortada** es un compromiso 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 para luego promediar lo que queda.

EJEMPLO 1.16 La producción de Bidri es una artesanía tradicional de India. Las artesanías Bidri (tazones, recipientes, etc.) se funden en una aleación que contiene principalmente zinc y algo de cobre. Considere las siguientes observaciones sobre el contenido de cobre (%) de una muestra de artefactos Bidri tomada del Museo Victoria y Albert de 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 1.17 es una gráfica de puntos de los datos. Una característica prominente es el valor atípico único en el extremo superior; la distribución está un tanto 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 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 $\bar{x}_{tr(7.7)} = 3.42$. El recorte en este caso elimina el valor extremo más grande y, por tanto, acerca la media recortada hacia la mediana.

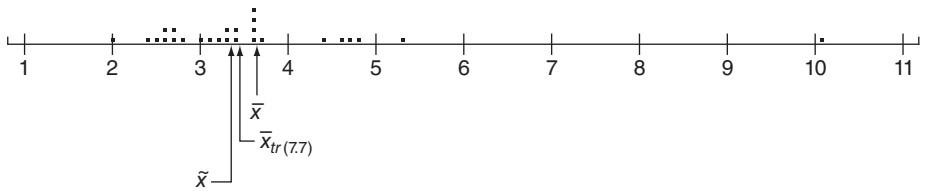


Figura 1.17 Gráfica de puntos del contenido de cobre para el ejemplo 1.16 ■

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 atípicos como la media ni tan insensible como la mediana. Si el porcentaje de recorte deseado es 100a% y na no es un entero, la media recortada debe ser calculada por interpolación. Por ejemplo, considere a = 0.10 para un porcentaje de recorte de 10% y $n = 26$ como en el ejemplo 1.16. Entonces $\bar{x}_{tr(10)}$ sería el promedio ponderado apropiado de la media recortada 7.7% calculada allí y la media recortada 11.5% que resulta de recortar tres observaciones de cada extremo.

41. Se eligió una muestra de $n = 10$ automóviles y cada una se sometió a una prueba de choque a 5 mph. Se denotó un auto sin daños visibles con S y uno con daños con F y 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. Después calcule $\bar{x}$ de esta muestra numéricamente codificada. ¿Cómo se compara $\bar{x}$ con x/n?

c. Suponga que se decide incluir 15 autos más en el experimento. ¿Cuántos de estos tendrían que ser S para obtener $x/n = 0.80$ 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 la mediana muestrales de las y_i con la media y la 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). Nuevamente 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. Se denotan con 100+ las duraciones de los dos componentes que continuaron funcionando después de 100 horas. Las observaciones muestrales resultantes fueron:

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

¿Cuáles de las medidas centrales 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 "censados a la derecha".]

1.4 Medidas de variabilidad

El reporte de una medida de centro sólo da información parcial sobre un conjunto o distribución de datos. Diferentes muestras o poblaciones pueden tener medidas idénticas de centro y aun así diferir una de otra en otras importantes maneras. La figura 1.18 muestra gráficas de puntos de tres muestras con las mismas media y mediana, aunque el grado de dispersión en 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 respecto a las otras dos en este aspecto.

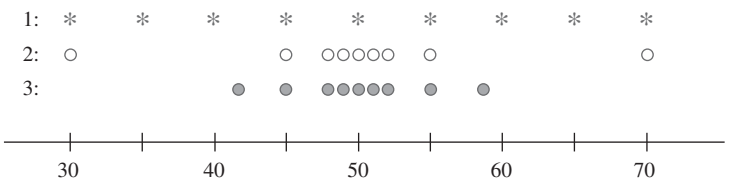


Figura 1.18 Muestras con medidas de centro idénticas pero diferentes cantidades de variabilidad

Medidas de variabilidad de datos muestrales

La medida más simple de variabilidad en una muestra es el **rango**, que es la diferencia entre el valor muestral más grande y el más pequeño. El valor del rango de la muestra 1 en la figura 1.18 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 valores restantes. Las muestras 1 y 2 en la figura 1.18 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 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 de la suma y el hecho de que $\Sigma\bar{x} = \bar{x} + \bar{x} + \cdots + \bar{x} = n\bar{x}$:

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

Existen maneras de 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 $\Sigma|x_i - \bar{x}|/n$. Debido a que la operación de valor absoluto conduce a un número de dificultades teóricas considere, 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 $\Sigma(x_i - \bar{x})^2/n$, 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:

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

Observe 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 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 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 $s = 3.0$ para una segunda muestra de autos de otro tipo, una desviación típica en esta muestra es aproximadamente 1.5 veces la de la primera, una indicación de más variabilidad en la segunda muestra.

EJEMPLO 1.17 El sitio web **www.fueleconomy.gov** contiene gran cantidad de información acerca de las características del combustible de varios vehículos. Además de las calificaciones de millaje de la Environmental Protection Agency (EPA), hay muchos usuarios de vehículos que han informado respecto a sus propios valores de eficiencia de combustible (mpg). Considere la siguiente muestra de $n = 11$ eficiencias para el Ford Focus 2009 equipado con transmisión

automática (para este modelo, la EPA informa de una calificación general de 27 mpg-24 mpg en ciudad y 33 mpg en carretera):

Automóvil	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	
1	27.3	−5.96	35.522	
2	27.9	−5.36	28.730	
3	32.9	−0.36	0.130	
4	35.2	1.94	3.764	
5	44.9	11.64	135.490	
6	39.9	6.64	44.090	
7	30.0	−3.26	10.628	
8	29.7	−3.56	12.674	
9	28.5	−4.76	22.658	
10	32.0	−1.26	1.588	
11	37.6	4.34	18.836	
	$\Sigma x_i = 365.9$	$\Sigma(x_i - \bar{x}) = 0.04$	$\Sigma(x_i - \bar{x})^2 = 314.110$	$\bar{x} = 33.26$

Debido al redondeo la suma de las desviaciones no da exactamente cero. El numerador de s^2 es $S_{xx} = 314.110$, por consiguiente

$$s^2 = \frac{S_{xx}}{n-1} = \frac{314.110}{11-1} = 31.41, \qquad s = 5.60$$

El tamaño de una desviación representativa de la media de la muestra 33.26 es de aproximadamente 5.6 mpg. *Nota:* De las nueve personas que también reportaron hábitos de conducción, sólo tres condujeron más de 80% en la autopista; apostamos a que puede adivinar los automóviles que conducían. Todavía no tenemos idea de por qué los 11 valores registrados exceden la cifra de la EPA, tal vez sólo los conductores con una realmente buena eficiencia de combustible comunican sus resultados. ■

Motivación para s^2

Para explicar el porqué del divisor $n - 1$ en s^2, observe 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). El valor de σ se puede interpretar como aproximadamente del tamaño de una desviación típica de μ dentro de toda la población de x valores. 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 respecto a la media poblacional (para la población, el divisor es N y no $N - 1$). En los capítulos 3 y 4 se presentan definiciones más generales de σ^2.

Así como se utilizará $\bar{x}$ para hacer inferencias sobre la media poblacional μ, se deberá definir la varianza muestral de modo que pueda ser utilizada para hacer inferencias sobre σ^2. Ahora observe que σ^2 implica desviaciones cuadradas 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 x_i 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 respecto a $\bar{x}$. Pero *las x_i tienden a acercarse más a su valor promedio $\bar{x}$ que el promedio poblacional* μ. Para compensar lo anterior 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 (en promedio se producen valores demasiado pequeños), mientras que si se divide entre el divisor un poco más pequeño $n - 1$ se corrige esta subestimación.

Es costumbre referirse a s^2 como si estuviera basada en $n - 1$ **grados de libertad** (gl). Esta terminología se deriva del hecho de que aunque s^2 está basada en las n cantidades $x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$, estas 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 $x_4 - \bar{x} = -4$, automáticamente $x_3 - \bar{x} = 2$, por lo que sólo tres de los cuatro valores de $x_i - \bar{x}$ son determinados libremente (3 grados de libertad).

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 que evita calcular las desviaciones. La fórmula implica a $(\sum x_i)^2$, sumar y luego elevar al cuadrado; y a $\sum x_i^2$, elevar al cuadrado y luego sumar.

Una expresión alternativa para el numerador de s^2 es

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

Demostración Debido a que $\bar{x} = \Sigma x_i / n$, $n(\bar{x})^2 = (\Sigma x_i)^2 / n$. 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 1.18 La luxación traumática de rodilla a menudo requiere cirugía para reparar los ligamentos rotos. Una medida de la recuperación es la amplitud de movimiento (medido como el ángulo formado cuando, a partir de la pierna estirada, la rodilla se dobla en la medida de lo posible). Los datos que figuran en el rango de movimiento posquirúrgico aparecen en el artículo **"Reconstruction of the Anterior and Posterior Cruciate Ligaments After Knee Dislocation" (*Amer. J. Sports Med*., 1999: 189-197):**

154 142 137 133 122 126 135 135 108 120 127 134 122

La suma de estas 13 muestras observadas es $\sum x_i = 1695$ y la suma de sus cuadrados es

$$\sum x_i^2 = (154)^2 + (142)^2 + \cdots + (122)^2 = 222.581$$

Por tanto, el numerador de la varianza muestral es

$$S_{xx} = \sum x_i^2 - \left[(\sum x_i)^2\right]/n = 222.581 - (1695)^2/13 = 1579.0769$$

de donde $s^2 = 1579.0769/12 = 131.59$ y $s = 11.47$. ■

Tanto la fórmula de la definición como la fórmula de cálculo para s^2 pueden ser sensibles al redondeo, por lo que en los cálculos intermedios se debe utilizar la mayor precisión decimal que sea posible.

Varias propiedades de s^2 pueden mejorar la comprensión y facilitar el cálculo.

PROPOSICIÓN

Sean $x_1, x_2, \ldots, x_n$ una muestra y c cualquier constante diferente de 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 otras palabras, el resultado 1 dice que si se suma (o se resta) una constante c de cada valor de dato, la varianza no cambia. Esto es intuitivo puesto que la adición o sustracción de c cambia la ubicación del conjunto de datos, pero deja inalteradas las distancias 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 en el resultado 1 que $\bar{y} = \bar{x} + c$ y en el resultado 2 que $\bar{y} = c\bar{x}$.

Gráficas de caja

Las gráficas de tallo y hojas y los 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 la naturaleza de cualquier alejamiento de la simetría, y 4) la identificación de las observaciones "atípicas" inusualmente alejadas del cuerpo principal de los datos. Puesto que 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 atípicos: la mediana y una medida de variabilidad llamada *distancia entre cuartos*.

DEFINICIÓN

Se ordenan las n observaciones de la más pequeña a la más grande y se separa la mitad más pequeña de la más grande; si n es impar se incluye la mediana en ambas mitades. 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 resistente a los valores atípicos es la distancia entre cuartos f_s, dada por

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

En general, la distancia entre 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 atípicos. Por tanto, es resistente a valores atípicos. Los cuartos son muy similares a los cuartiles y la cuarta extensión es similar al *rango intercuartil*, la diferencia entre los cuartiles superiores e inferiores. Pero los cuartiles son un poco más enfadosos que los cuartos para calcular a mano, y existen diferentes maneras razonables para calcular los cuartiles (así los valores pueden variar de un programa informático a otro).

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 coloca un rectángulo sobre una escala de medición horizontal; el lado izquierdo del rectángulo está arriba 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 adentro del rectángulo en la ubicación de la mediana; la posición del símbolo de la mediana respecto a los dos lados da información sobre asimetría en el 50% medio de los datos. Por