1.17 (Optativo). La impresión proporcionada por los gráficos temporales depende de la escala que se emplea en los ejes de abcisas y de ordenadas. Si alargas el eje de ordenadas y comprimes el eje de abcisas, los cambios parecen ser más rápidos. Sin embargo, si se comprime el eje de ordenadas y se alarga el eje de abcisas los cambios parecen más lentos. Representa dos gráficos temporales más con los datos del ejemplo 1.5, uno que haga que la tasa de mortalidad por cáncer parezca que crece muy rápidamente y uno que muestre sólo un incremento suave. La intención de este ejercicio es ponerte en guardia. Cuando observes un gráfico temporal fíjate muy bien en las escalas de los ejes.

La tabla 1.2 presenta datos sobre los Estados europeos. El estudio de un conjunto de datos con muchas variables empieza con el análisis individual de cada variable. Los ejercicios 1.18, 1.19 y 1.20 hacen referencia a estos datos.

1.18. Dibuja un gráfico de la distribución del número de periódicos por cada 1.000 habitantes en los diversos Estados europeos. Describe brevemente el aspecto general de la distribución y las posibles observaciones atípicas.

1.19. Dibuja un gráfico que muestre la distribución del gasto público en educación en los distintos Estados. ¿Cuál es el aspecto general de esta distribución? ¿Se observa alguna observación atípica o alguna desviación?

1.20. Dibuja un gráfico de la distribución de la renta per cápita en los distintos Estados. ¿Tiene el gráfico un aspecto general bien definido? ¿Existe alguna observación atípica o alguna otra desviación importante?

1.3 Descripción de las distribuciones con números

¿Qué edad tenían los presidentes de EE UU al inicio de su mandato? Bill Clinton tenía 46 años, ¿era muy joven? La tabla 1.3 da las edades de todos los presidentes de EE UU al inicio de su mandato. Como muestra el histograma de la figura 1.10, hay una variación importante en las edades de los presidentes. Teddy Roosevelt fue, con 42 años, el más joven y Ronald Reagan, con 69 años, el mayor. La distribución es aproximadamente simétrica. El gráfico muestra que la edad de un presidente típico de los EE UU al inicio de su mandato es de aproximadamente 55 años, ya que 55 se encuentra cerca del centro del histograma.

Hemos ofrecido una breve descripción de la distribución que incluía su *forma* (aproximadamente simétrica), un número para describir su *centro* (aproximadamente 55), y otros números para describir su *dispersión* (de 42 a 69). La forma, el centro y la dispersión proporcionan una buena descripción del aspecto general de cualquier

Tabla 1.3. Edades de los presidentes de EE UU al inicio de su mandato.

Presidente	Edad	Presidente	Edad	Presidente	Edad
Washington	57	Buchanan	65	Harding	55
J. Adams	61	Lincoln	52	Coolidge	51
Jefferson	57	A. Johnson	56	Hoover	54
Madison	57	Grant	46	F. D. Roosevelt	51
Monroe	58	Hayes	54	Truman	60
J. Q. Adams	57	Garfield	49	Eisenhower	61
Jackson	61	Arthur	51	Kennedy	43
Van Buren	54	Cleveland	47	L. Johnson	55
W. H. Harrison	68	B. Harrison	55	Nixon	56
Tyler	51	Cleveland	55	Ford	61
Polk	49	McKinley	54	Carter	52
Taylor	64	T. Roosevelt	42	Reagan	69
Fillmore	50	Taft	51	Bush	64
Pierce	48	Wilson	56	Clinton	46

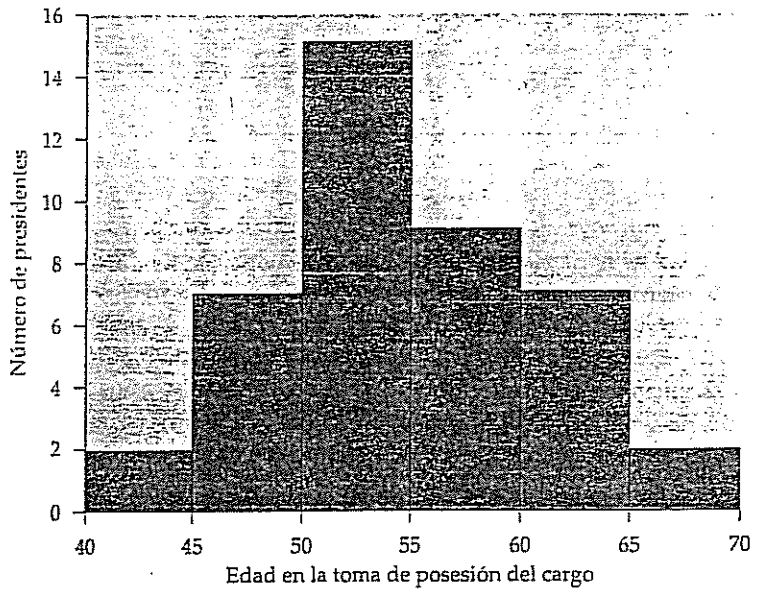

Figura 1.10. Distribución de las edades de los presidentes de EE UU en el momento de tomar posesión de su cargo. Datos de la tabla 1.3.

distribución de una variable cuantitativa. Ahora aprenderemos a medir el centro y la dispersión de una distribución. Podemos calcular estas medidas numéricas para cualquier variable cuantitativa. Pero para interpretar las medidas de centro y de dispersión, y para escoger entre las distintas medidas que aprenderemos, tienes que tener en cuen-

ta la forma de la distribución y el significado de los datos. Los números, igual que los gráficos, son ayudas para comprender, no son por sí mismos "la respuesta".

1.3.1 Una medida de centro: la media

La descripción de una distribución casi siempre incluye una medida de su centro o de su promedio. La medida de centro más común es la *media* aritmética.

> LA MEDIA $\bar{x}$
>
> Si n observaciones se denotan como $x_1, x_2, \ldots, x_n$, su media es
>
> $$\bar{x} = \frac{(x_1 + x_2 + \ldots + x_n)}{n}$$
>
> o, de una forma más compacta,
>
> $$\bar{x} = \frac{1}{n}\sum x_i$$

La Σ (letra griega sigma mayúscula) en la fórmula de la media significa "suma de todos los elementos". Los subíndices de las observaciones x_i son una forma de distinguirlas. No indican necesariamente ni orden ni ninguna característica especial de los datos. La barra sobre la x simboliza la media de todos los valores de x. Nos referimos a $\bar{x}$ como a "x barra". Esta notación es muy común. Cuando utilizamos los símbolos $\bar{x}$ o $\bar{y}$ siempre nos referimos a una media aritmética.

EJEMPLO 1.6

La tabla 1.3 da las edades de los 42 presidentes de los EE UU cuando accedieron al cargo. La media de estos valores es

$$\bar{x} = \frac{(x_1 + x_2 + \ldots + x_n)}{n} =$$
$$= \frac{57 + 61 + 57 + \ldots + 46}{42} =$$
$$= \frac{2.303}{42} = 54{,}8 \text{ años}$$

En la práctica, puedes hacer los cálculos con la calculadora. No es necesario que los hagas a mano, pero sí deberías entender lo que hace la calculadora. ■

EJEMPLO 1.7

Un estudio realizado en Suiza examinó el número de histerectomías (extirpación del útero) realizadas durante un año por un grupo de médicos. He aquí los datos de una muestra de 15 cirujanos.

27 50 33 25 86 25 85 31 37 44 20 36 59 34 28

El diagrama de tallos muestra que la distribución es asimétrica hacia la derecha y que hay dos observaciones atípicas en el lado de los valores mayores.

2	05578
3	13467
4	4
5	09
6	
7	
8	56

Utiliza la calculadora para ver que la media de las operaciones realizadas por estos 15 médicos es $\bar{x} = 41{,}3$. Luego, fíjate en que sólo 5 de los 15 médicos realizaron un número de histerectomías mayor que la media. Esto se debe a que las dos observaciones atípicas (85 y 86) hacen aumentar el valor de la media. Comprueba que la media de las restantes 13 observaciones es 34,5. La media es el promedio de los valores, pero *no* es el número de operaciones realizadas por un cirujano típico. ■

El ejemplo 1.7 ilustra un hecho importante sobre la media como medida de centro: la media es sensible a la influencia de unas pocas observaciones extremas. Pueden ser observaciones atípicas, pero una distribución asimétrica que no tenga observaciones atípicas también desplazará la media hacia la cola más larga.

EJERCICIOS

1.21. He aquí los resultados de 18 estudiantes universitarias de primer curso en la prueba SSHA (*Survey of Study Habits and Attitudes*) sobre hábitos de estudio y actitud de los estudiantes:

154	109	137	115	152	140	154	178	101
103	126	126	137	165	165	129	200	148

(a) Halla sin calculadora la media de estos datos utilizando la fórmula. Ahora, calcula la media con la ayuda de la calculadora. Comprueba que obtienes el mismo resultado.

(b) El diagrama de tallos del ejercicio 1.7 sugiere que una puntuación de 200 es una observación atípica. Utiliza la calculadora para hallar la media prescindiendo de dicho valor. Describe brevemente cómo la observación atípica modifica la media.

1.3.2 Una medida de centro: la mediana

La media no es la única forma de describir el centro de una distribución. Otra posibilidad es utilizar "el valor central" de un histograma o de un diagrama de tallos. Es decir, halla un número tal que la mitad de las observaciones sean menores y la otra mitad mayores. Dicho número es la *mediana* de una distribución. Para abreviar, a la mediana la llamaremos M. Aunque la idea de la mediana como el punto medio de una distribución es sencilla, necesitamos una regla precisa para llevar a la práctica esta idea. La regla aparece en el siguiente recuadro.

LA MEDIANA M

Para hallar la mediana de una distribución:

1. Ordena todas las observaciones de la mínima a la máxima.
2. Si el número de observaciones n es impar la mediana M es la observación central de la lista ordenada. Halla la posición de la mediana contando $(n + 1)/2$ observaciones desde el comienzo de la lista.
3. Si el número de observaciones n es par, entonces la mediana M es la media de las dos observaciones centrales de la lista ordenada. La posición de la mediana se halla, otra vez, contando $(n + 1)/2$ desde el comienzo de la lista.

Las calculadoras de bolsillo no suelen tener una función para el cálculo de la mediana. Por tanto, tendrás que calcularla paso a paso a no ser que utilices un ordenador o una calculadora con funciones estadísticas avanzadas. Todos los programas estadísticos calculan las medianas. Como ordenar muchas observaciones lleva tiempo, utiliza un ordenador si tienes muchas observaciones. Cuando calcules la mediana sin la calculadora, asegúrate de no olvidar ninguna observación, especialmente si algunas tienen valores repetidos. Asegúrate, también, de colocar todas las observaciones en orden creciente antes de hallar la mediana. La observación central de las observaciones colocadas según el orden en que fueron obtenidas no tiene ningún significado. He aquí un ejemplo que muestra cómo se aplica la regla del cálculo de la mediana, para un número par y un número impar de observaciones.

EJEMPLO 1.8

Para hallar la mediana de las histerectomías realizadas por los 15 médicos del ejemplo 1.7, en primer lugar ordena las observaciones en orden creciente:

20 25 25 27 28 31 33 34 36 37 44 50 59 85 86

Hay un número impar de observaciones, por tanto, hay una observación central. Es la mediana. El número 34 marcado en negrita en la lista es la observación central, ya que tiene 7 observaciones a su izquierda y 7 a su derecha. Por consiguiente, la mediana es $M = 34$.

Es más rápido utilizar la regla para localizar la mediana de la lista. Como $n = 15$, la posición de la mediana es

$$\frac{n+1}{2} = \frac{16}{2} = 8$$

Es decir, la mediana es la octava observación de la lista ordenada.

El estudio suizo también analizó una muestra de las histerectomías realizadas por un grupo de 10 cirujanos mujeres. El número de histerectomías realizadas por las cirujanas (dispuestos en orden) fueron

5 7 10 14 18 19 25 29 31 33

Aquí $n = 10$ es par. No hay una observación central, hay un par central. Son los valores 18 y 19 en negrita de la lista. Ambos valores tienen cuatro observaciones a la izquierda y cuatro a la derecha. La mediana se encuentra a medio camino de estas observaciones:

$$M = \frac{18 + 19}{2} = 18{,}5$$

La regla para localizar la mediana en la lista da

$$\frac{n+1}{2} = \frac{11}{2} = 5{,}5$$

La posición 5,5 significa "a medio camino entre las observaciones quinta y sexta de la lista ordenada". Esto concuerda con lo que hemos observado a simple vista.

La cirujana típica realizó muchas menos histerectomías que el cirujano típico. Ésta fue una de las conclusiones importantes del estudio. ■

1.3.3 Comparación entre la media y la mediana

Los ejemplos 1.7 y 1.8 muestran una diferencia importante entre la media y la mediana. El ejemplo 1.7 muestra cómo las dos observaciones atípicas tiran de la media. Pero las observaciones atípicas no tienen ningún efecto sobre la mediana. Las observaciones atípicas tan sólo son dos valores más de entre la mitad de valores mayores. La mediana permanecería igual incluso si un cirujano hubiera realizado 1.000 operaciones. De todas formas, esto no significa que como a la mediana no la influyen unas pocas observaciones extremas, siempre sea mejor que la media. La media y la mediana son dos maneras distintas de medir el centro y ambas son útiles. Utiliza la mediana si quieres el número de histerectomías realizadas por un cirujano típico. Utiliza la media si también estás interesado en el número total de operaciones realizadas por todos los cirujanos. El número total de operaciones es n veces la media donde n es el número de cirujanos. El total de operaciones no se puede calcular a partir de la mediana.

La media y la mediana de una distribución simétrica se encuentran muy cerca. Si la distribución es *exactamente* simétrica, la media y la mediana son exactamente iguales. En una distribución asimétrica, la media queda desplazada hacia la cola más larga. Por ejemplo, la distribución del precio de las viviendas es asimétrica hacia la derecha. Existen muchas viviendas de precio moderado y unas cuantas que son muy caras. Las pocas viviendas caras tiran de la media y, sin embargo, no afectan a la mediana. Por ejemplo, el precio medio de todas las casas vendidas en 1993 fue de 139.400 dólares. En cambio, el precio mediano fue de 117.000 dólares. En los informes sobre los precios de las viviendas, sobre los ingresos y sobre otras distribuciones muy asimétricas normalmente se calcula la mediana ("el valor típico") en lugar de la media ("el valor promedio").

EJERCICIOS

1.22. He aquí el número de carreras de béisbol realizadas cada año por Babe Ruth durante los 15 años que jugó con los New York Yankees:

54 59 35 41 46 25 47 60 54 46 49 46 41 34 22

y por Roger Maris durante 10 años en la liga americana:

13 23 26 16 33 61 28 39 14 8

Halla la mediana del número de carreras realizadas por cada uno de estos deportistas.

1.23. En el ejercicio 1.21 hallamos la media de los resultados en la prueba SSHA de 18 estudiantes universitarias de primer curso. Ahora, calcula la mediana de estos resultados. ¿La mediana es mayor o menor que la media? Explica por qué ocurre de esta manera.

1.24. En el año 1994, una pequeña empresa pagó 2.200.000 pesetas a cada uno de sus cinco administrativos y 5.000.000 de pesetas a cada uno de sus dos contables. El gerente de la empresa recibió 27.000.000 de pesetas. ¿Cuál es el salario medio pagado por esta empresa? ¿Cuántos empleados ganan menos que la media? ¿Cuál es el valor del salario mediano?

1.25. La media y la mediana de los sueldos pagados a los mejores jugadores de la liga americana de béisbol en 1993 fue de 490.000 dólares y de 1.160.000 dólares. ¿Cuál de estas cifras crees que es la media y cuál la mediana? Justifica tu respuesta.

1.3.4 Una medida de dispersión: los cuartiles

La media y la mediana proporcionan dos medidas distintas del centro de una distribución. Caracterizar una distribución sólo con una medida de su centro puede ser engañoso. Dos provincias con el mismo ingreso mediano por hogar son muy distintas si una de ellas tiene extremos de pobreza y de riqueza, mientras que la otra tiene poca variación entre familias. Un lote de medicinas con una concentración media adecuada en su componente activo puede ser muy peligroso si hay comprimidos con contenidos del componente activo muy elevados y otros con contenidos muy bajos. Estamos interesados en la *dispersión* o *variabilidad* de los ingresos o de las concentraciones del componente activo en las medicinas, además de estarlo en sus centros.

La descripción numérica útil más simple de una distribución consiste en una medida de centro y una medida de dispersión.

Recorrido

Una manera de medir la dispersión es calcular el ***recorrido***, es decir, la diferencia entre las observaciones máxima y mínima.

EJEMPLO 1.9

El ejemplo 1.7 proporciona datos sobre el número de histerectomías realizadas durante un año por un grupo de 15 cirujanos. El número mínimo fue 20 y el máximo 86. Por tanto, el recorrido fue

$$\text{recorrido} = 86 - 20 = 66 \quad \blacksquare$$

Cuartiles

El recorrido muestra la variación total de los datos. Depende sólo de las observaciones máxima y mínima, que podrían ser observaciones atípicas. Podríamos mejorar nuestra descripción de la dispersión fijándonos también en la dispersión del 50% de los valores centrales de nuestros datos. Los ***cuartiles*** determinan entre qué valores se encuentra la mitad central de las observaciones. Cuenta en orden creciente en la lista ordenada de observaciones, empezando por la menor. El *primer cuartil* se sitúa en el primer 25% de las observaciones. El *tercer cuartil* se sitúa en el primer 75%. En otras palabras, el primer cuartil es mayor que el 25% de las observaciones y el tercer cuartil es mayor que el 75%. El segundo cuartil es la mediana, que es mayor que el 50% de las observaciones. Esta es la idea de los cuartiles. Necesitamos una regla para concretar esta idea. La regla para calcular los cuartiles utiliza la regla de la mediana.

He aquí un ejemplo que muestra cómo se aplica la regla de los cuartiles para un número impar y un número par de observaciones.

EJEMPLO 1.10

El número de histerectomías realizadas por nuestra muestra de 15 cirujanos es (colocados en orden):

20 25 25 27 28 31 33 34 36 37 44 50 59 85 86

CUARTILES Q_1 Y Q_3

Para calcular los **cuartiles**:

1. Ordena las observaciones en orden creciente y localiza la mediana M en la lista ordenada de observaciones.
2. El primer cuartil Q_1 es la mediana de las observaciones situadas a la izquierda de la mediana de la totalidad.
3. El tercer cuartil Q_3 es la mediana de las observaciones situadas a la derecha de la mediana de la totalidad.

Hay un número de observaciones impar, por tanto, la mediana es el valor central, es decir, el número 34 de la lista. El primer cuartil es la mediana de las 7 observaciones situadas a la izquierda de la mediana. Es la cuarta de estas 7 observaciones, por tanto, $Q_1 = 27$. Si quieres, puedes utilizar la regla de la mediana con $n = 7$:

$$\frac{n+1}{2} = \frac{7+1}{2} = 4$$

El tercer cuartil es la mediana de las 7 observaciones situadas a la derecha de la mediana, $Q_3 = 50$. La mediana de la totalidad se deja fuera de los cálculos cuando hay un número impar de observaciones.

Para las 10 cirujanas los datos son (otra vez ordenados en orden creciente)

5 7 10 14 18 | 19 25 29 31 33

Hay un número par de observaciones, por tanto, la mediana se encuentra entre el par central. Se sitúa entre las observaciones quinta y sexta, que señalamos con el símbolo |. El primer cuartil es la mediana de las primeras cinco observaciones, ya que éstas son las observaciones situadas a la izquierda de la mediana. Comprueba que $Q_1 = 10$ y que $Q_3 = 29$. Cuando el número de observaciones es par, todas las observaciones entran en el cálculo de los cuartiles. ■

Ve con cuidado cuando diversas observaciones toman el mismo valor numérico. Escribe todas las observaciones y aplica las reglas como si todos los valores fueran distintos. Por ejemplo, la mediana de

4 7 7 7 8 9 9

es $M = 7$, ya que el número 7 marcado en negrita es la observación central de la lista. El primer cuartil es la mediana de las tres observaciones situadas a la izquierda del 7, que son 4, 7, 7. Por tanto, el primer cuartil también es 7. El tercer cuartil es el número 9.

Algunos programas estadísticos utilizan una regla un poco distinta para hallar los cuartiles, por lo que los resultados del ordenador pueden ser algo distintos a los que calcules a mano. Pero no te preocupes por eso, ya que las diferencias serán siempre demasiado pequeñas para ser importantes.

1.3.5 Los cinco números resumen y los diagramas de caja

Una manera conveniente de describir el centro y la dispersión de un conjunto de datos es dar la mediana para medir el centro, y los cuartiles y las observaciones individuales mínima y máxima para indicar la dispersión.

Estos cinco números proporcionan una descripción razonablemente completa del centro y la dispersión. Los cinco números resumen del ejemplo 1.10 son

20 27 34 50 86

para los cirujanos y

5 10 18.5 29 33

para las cirujanas.

> **LOS CINCO NÚMEROS RESUMEN**
>
> Los **cinco números resumen** de un conjunto de datos consisten en la observación mínima, el primer cuartil, la mediana, el tercer cuartil y la observación máxima, escritos en orden de menor a mayor. De forma simbólica son
>
> Mínima Q_1 M Q_3 Máxima

Diagrama de caja

Los cinco números resumen de una distribución nos llevan a un nuevo diagrama, el ***diagrama de caja***. La figura 1.11 muestra los diagramas de caja de los cirujanos suizos. Los lados inferior y superior de la caja van del primer al tercer cuartil. Por tanto, la altura de la caja es la amplitud del 50% de los datos centrales. El segmento del interior de la caja indica la mediana. Los extremos de los segmentos perpendiculares a los lados superior e inferior indican la posición de los valores máximo y mínimo, respectivamente. Podemos ver, de forma rápida, que las cirujanas hacen en promedio menos histerectomías que los cirujanos, y también que hay menos variación entre las cirujanas.

Puedes dibujar los diagramas de caja en posición horizontal o vertical. Asegúrate de incluir siempre una escala en el dibujo. Cuando mires un diagrama de caja, localiza en primer lugar la mediana que sitúa el centro de la distribución. Luego, fíjate en la dispersión. Los cuartiles muestran la dispersión del 50% de los datos centrales, los extremos del diagrama de caja (observaciones máxima y mínima) muestran la dispersión de todos los datos. La situación relativa de los lados de la caja, y de los extremos de los segmentos exteriores, respecto a la mediana, dan una indicación de la simetría o de la asimetría de la distribución. En una distribución simétrica, el primer y el tercer cuartil están aproximadamente a la misma distancia de la mediana. En la mayoría de las distribuciones asimétricas hacia la derecha, en cambio, el tercer cuartil estará situado mucho más a la derecha de la mediana que el primer cuartil a su izquierda. Los extremos se comportan de la misma manera; pero recuerda que sólo son observaciones individuales y puede ser que digan poco sobre la distribución como un todo. Como los diagramas de caja muestran menos detalles que los histogramas o los diagramas de tallos, es mejor utilizarlos para la comparación de más de una distribución en un mismo gráfico, como el de la figura 1.11.

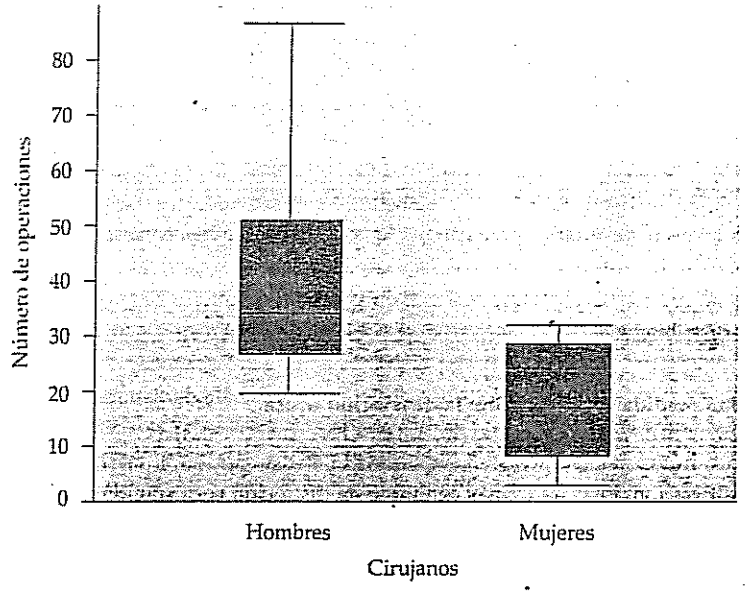

Figura 1.11. Diagramas de caja en un mismo gráfico para comparar el número de histerectomías realizadas por los cirujanos y las cirujanas.

EJERCICIOS

1.26. El ejercicio 1.22 da el número de carreras de béisbol realizadas cada año por Babe Ruth y Roger Maris.

(a) Halla los cinco números resumen de cada jugador.

(b) Compara ambos diagramas de caja en un mismo gráfico. ¿Qué puedes concluir?

1.27. Fíjate ahora en los datos presentados en la tabla 1.3. En el ejemplo 1.6 vimos que la edad media a la que se alcanzaba la presidencia en EE UU era 54.8 años.

(a) Observando la forma del histograma presentado en la figura 1.10, ¿crees que la mediana será mucho menor que la media, similar a la media o muy superior?

(b) Calcula los cinco números resumen y comprueba que la mediana está donde tú esperabas hallarla.

(c) ¿Cuál es el recorrido del 50% de las observaciones centrales?

1.28. La tabla 1.2 contiene datos sobre los Estados europeos. Queremos comparar las distribuciones del número de periódicos y de televisores por cada 1.000 habitantes. Hemos introducido estos datos en el ordenador con los nombres PERIOD (para los periódicos) y TELEV (para los televisores). He aquí los resultados del programa estadístico utilizado (Minitab) que nos dan los cinco números resumen junto con otra información (otros programas ofrecen resultados similares).

	N	N*	MEAN	MEDIAN	TRMEAN	STDEV	SEMEAN	MIN	MAX	Q1	Q3
PERIOD	40	2	269.5	241.5	262.4	176.0	27.8	27.0	653.0	121.3	383.0
TELEV	40	2	380.4	377.5	374.4	138.3	21.9	89.0	745.0	297.2	467.5

Utiliza estos resultados para dibujar los diagramas de caja correspondientes a la distribución del número de periódicos y de televisores por cada 1.000 habitantes en un mismo gráfico. Expresa brevemente en palabras las diferencias y similitudes de ambas distribuciones.

1.3.6 Una medida de dispersión: la desviación típica

Los cinco números resumen son, seguramente, la descripción numérica más útil de una distribución, pero no la más común. Esta distinción corresponde a la combinación de la media para medir el centro y la *desviación típica* para medir la dispersión. La desviación típica mide la dispersión de las observaciones respecto a la media.

En la práctica, utiliza una calculadora o un ordenador para calcular la desviación típica. De todas maneras, calcular algunos casos paso a paso te ayudará a comprender el cálculo de la varianza y la desviación típica. He aquí un ejemplo.

LA DESVIACIÓN TÍPICA s

La varianza s^2 de un conjunto de observaciones es el promedio de los cuadrados de las desviaciones de las observaciones respecto a su media. Con símbolos, la varianza de n observaciones, $x_1, x_2, \ldots, x_n$ es

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n - 1}$$

o de una forma más simple,

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

La **desviación típica** es la raíz cuadrada positiva de la varianza s^2:

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

EJEMPLO 1.11

El nivel metabólico de una persona es el ritmo al que el cuerpo consume energía. El nivel metabólico es importante en los estudios de dietética. He aquí los niveles metabólicos de 7 hombres que tomaron parte en un estudio de dietética (las unidades son calorías por 24 horas. Las calorías también se utilizan para describir el contenido energético de los alimentos).

1.792 1.666 1.362 1.614 1.460 1.867 1.439

Los investigadores calcularon la $\overline{x}$ y la s de estos hombres.

En primer lugar, halla la media:

$$\overline{x} = \frac{1.792 + 1.666 + 1.362 + 1.614 + 1.460 + 1.867 + 1.439}{7} =$$

$$= \frac{11.200}{7} = 1.600$$

Para ver claramente la naturaleza de la varianza, empieza con una tabla de las desviaciones de las observaciones respecto a esta media.

Observaciones x_i	Desviaciones $x_i - \bar{x}$	Desviaciones al cuadrado $(x_i - \bar{x})^2$
1.792	1.792 – 1.600 = 192	192^2 = 36.864
1.666	1.666 – 1.600 = 66	66^2 = 4.356
1.362	1.362 – 1.600 = –238	$(-238)^2$ = 56.644
1.614	1.614 – 1.600 = 14	14^2 = 196
1.460	1.460 – 1.600 = –140	$(-140)^2$ = 19.600
1.867	1.867 – 1.600 = 267	267^2 = 71.289
1.439	1.439 – 1.600 = –161	$(-161)^2$ = 25.921
	suma = 0	suma = 214.870

La varianza es la suma de las desviaciones al cuadrado dividido por el número de observaciones menos uno.

$$s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2 = \frac{1}{2}\,(214.870) = 35.811.67$$

La desviación típica es la raíz cuadrada positiva de la varianza:

$$s = \sqrt{35.811.67} = 189.24 \qquad \blacksquare$$

La figura 1.12 muestra los datos del ejemplo 1.11 como puntos sobre una línea horizontal numerada. La media se ha simbolizado con un asterisco (*). Las flechas indican las desviaciones de dos observaciones respecto a la media. Estas desviaciones muestran la dispersión de los datos respecto a su media. Algunas desviaciones serán positivas y otras negativas, ya que unas quedan a la derecha de la media y otras a su izquierda. De hecho, *la suma de todas las desviaciones respecto a la media es siempre cero*. Comprueba que esto es cierto en el ejemplo 1.11. Por tanto, no podemos sumar simplemente las desviaciones para hallar una medida conjunta de dispersión. Una manera de solventar este inconveniente es elevar al cuadrado las desviaciones individuales y sumar estos cuadrados. La varianza s^2 es el promedio de estas desviaciones al cuadrado. La varianza es grande si las observaciones están muy dispersas respecto a la media, y es pequeña si todas las observaciones se sitúan cerca de la media.

Como la varianza exige elevar al cuadrado las desviaciones, no tiene las mismas unidades de medida que las observaciones originales. Por ejemplo, las longitudes medidas en centímetros tienen una varianza expresada en centímetros al cuadrado. Si obtenemos la raíz cuadrada el problema queda solucionado. La desviación típica s mide la dispersión de los datos respecto a la media en la escala original.

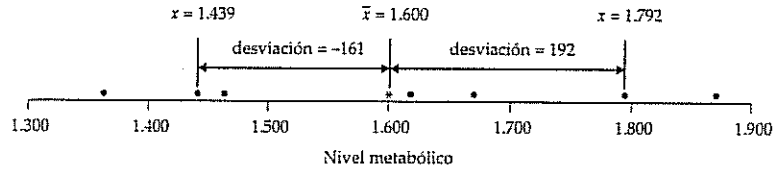

Figura 1.12. Niveles metabólicos de siete hombres, con la media (*) y las desviaciones de dos observaciones respecto a la media.

Grados de libertad

La varianza es el promedio de las desviaciones al cuadrado de las observaciones respecto a su media. ¿Por qué calculamos el promedio dividiendo por $n - 1$, en vez de dividir por n? Como la suma de las desviaciones siempre es cero, la última desviación se puede hallar cuando se conocen las otras $n - 1$. Por tanto, no estamos calculando el promedio de n números independientes. Sólo $n - 1$ de las desviaciones al cuadrado pueden variar libremente, y nosotros promediamos dividiendo el total por $n - 1$. Al número $n - 1$ se le denomina ***grados de libertad*** de la varianza o de la desviación típica. Muchas calculadoras ofrecen la posibilidad de dividir por n o por $n - 1$, por consiguiente, asegúrate de dividir por $n - 1$.

Dejando la aritmética a una calculadora o a un ordenador podremos concentrarnos en lo que hacemos y en por qué lo hacemos. Lo que estamos haciendo es medir la dispersión. He aquí las propiedades básicas de la desviación típica s como medida de dispersión.

PROPIEDADES DE LA DESVIACIÓN TÍPICA

- s mide la dispersión respecto a la media. Debe emplearse sólo cuando se escoge la media como medida de centro.
- $s = 0$ sólo cuando *no hay dispersión*. Esto ocurre únicamente cuando todas las observaciones tienen el mismo valor. De lo contrario $s > 0$. A medida que las observaciones están más dispersas respecto a su media, s se hace mayor.
- s, al igual que $\bar{x}$, está fuertemente influenciada por las observaciones extremas. Unas pocas observaciones atípicas pueden hacer que s sea muy grande.

Una distribución asimétrica con unas pocas observaciones en la cola larga de la distribución tendrá una desviación típica grande. En tal caso, el número s no proporciona una información demasiado útil. Como en una distribución muy asimétrica la dispersión de cada una de las colas es muy distinta, es imposible describir bien la dispersión con un solo número. Los cinco números resumen, con los dos cuartiles y los dos valores extremos, proporcionan una información mejor. *Es preferible utilizar los cinco números resumen en lugar de la media y la desviación típica para describir una distribución asimétrica. Utiliza $\bar{x}$ y s sólo para distribuciones razonablemente simétricas.*

Puede ser que creas que la importancia de la desviación típica no está suficientemente justificada. La desviación típica es algo complicada y, además, no ofrece una buena descripción de las distribuciones asimétricas. En la próxima sección veremos que la desviación típica es la medida natural de la dispersión para una clase de distribuciones simétricas: las distribuciones normales. La utilidad de muchos procedimientos estadísticos está ligada a la existencia de distribuciones con formas determinadas. Esto es especialmente cierto en el caso de la desviación típica.

Recuerda que la mejor visión global de una distribución la da un gráfico. Las medidas numéricas de centro y de dispersión reflejan características concretas de una distribución, pero no describen completamente su forma. Los resúmenes numéricos no detectan, por ejemplo, la presencia de múltiples picos o de espacios vacíos. El ejercicio 1.31 da un ejemplo de una distribución para la cual los resúmenes numéricos son engañosos. REPRESENTA SIEMPRE TUS DATOS GRÁFICAMENTE.

EJERCICIOS

1.29. La concentración de determinadas sustancias en la sangre influye en la salud de las personas. He aquí las mediciones del nivel de fosfato en la sangre de un paciente que realizó seis visitas consecutivas a una clínica, expresadas en miligramos de fosfato por decilitro de sangre.

5,6 5,2 4,6 4,9 5,7 6,4

Un gráfico con sólo 6 observaciones da poca información, por tanto, pasamos a calcular la media y la desviación típica.

(a) Halla la media a partir de su definición. Es decir, halla la suma de las 6 observaciones y divide por 6.

(b) Halla la desviación típica a partir de su definición. Es decir, calcula la desviación de cada observación respecto a su media y eleva estas desviaciones al cuadrado. Luego, calcula la varianza y la desviación típica. El ejemplo 1.11 ilustra este método.

(c) Ahora introduce los datos en la calculadora y halla la media y la desviación típica. ¿Has obtenido los mismos resultados que en los cálculos hechos a mano?

1.30. El número de histerectomías realizadas por el grupo de cirujanos durante un año descritas en el ejemplo 1.7 eran:

27 50 33 25 86 25 85 31 37 44 20 36 59 34 28

El diagrama de tallos muestra que las dos observaciones mayores son observaciones atípicas.

(a) El número medio de operaciones es $\bar{x} = 41{,}3$. Calcula a mano la desviación típica s. Es decir, calcula las desviaciones de cada observación respecto a su media y elévalas al cuadrado. Halla la varianza s^2 y la desviación típica s. Sigue el modelo del ejemplo 1.11.

(b) Introduce ahora los datos en la calculadora y verifica los resultados. Calcula $\bar{x}$ y s para las trece observaciones que nos quedan una vez eliminadas las observaciones atípicas. ¿Cómo afectan las observaciones atípicas a los valores $\bar{x}$ y s?

1.31. El ejercicio 1.28 da los resúmenes numéricos del número de periódicos y de televisores por cada 1.000 habitantes en los distintos Estados. Estos resúmenes numéricos (y los diagramas de caja derivados de ellos) *no* muestran una de las características más importantes de las distribuciones. Dibuja un diagrama de tallos con el número de televisores de la tabla 1.2. ¿Cuál es la forma de la distribución de esta variable? Recuerda empezar siempre el análisis de los datos con un gráfico –los resúmenes numéricos no son descripciones completas–.

RESUMEN

Un resumen numérico de una distribución tiene que dar su **centro** y su **dispersión** o **variabilidad**.

La **media** $\bar{x}$ y la **mediana** M describen el centro de una distribución de maneras distintas. La media es el promedio aritmético de las observaciones; la mediana es el punto medio de los valores.

Cuando utilices la mediana para indicar el centro de la distribución, describe su dispersión dando los **cuartiles**. El **primer cuartil** Q_1 tiene el 25% de las observaciones a su izquierda, el **tercer cuartil** Q_3 tiene el 75% de las observaciones a su izquierda.

Los **cinco números resumen** consisten en la mediana, los cuartiles y las observaciones extremas máxima y mínima, y proporcionan una descripción rápida de una