DAVID S. MOORE
Purdue University

ESTADÍSTICA APLICADA BÁSICA

Traducción y adaptación de
Jordi Comas
Universitat Pompeu Fabra

Antoni Bosch editor

Publicado por Antoni Bosch, editor
Manuel Girona, 61 – 08034 Barcelona
Tel. (34) 932 05 26 06 – Fax (34) 932 80 48 02
e-mail: antonibosch.editor@bcn.servicom.es
http://www.seker.es/insite/antonibosch

Título original de la obra:
The Basic Practice of Statistics

Impreso en España
Printed in Spain

ISBN: 84-85855-80-9
Depósito legal: B-49.187-1998

Diseño de la cubierta: Compañía de Diseño

Fotocomposición: Alemany, S.C.C.L.
Impresión: LiberDúplex

Impreso en papel reciclado

ÍNDICE DE CONTENIDO

terio no se pueden (al menos hasta ahora) automatizar, pero te sirven de guía para indicar al ordenador lo que tiene que hacer y para interpretar los resultados obtenidos. Este libro trata de explicar las ideas más importantes de la estadística, no sólo enseña métodos estadísticos. Algunos ejemplos de grandes ideas que encontrarás en este libro (una de cada uno de los tres campos de estudio de la estadística aplicada) son: "representa siempre tus datos gráficamente", "los experimentos comparativos aleatorizados" y "la significación estadística".

Los siguientes principios deben, pues, guiar tu estudio de la estadística:

- Intenta comprender qué dicen los datos en un determinado contexto. Todos los métodos que aprenderás sólo son herramientas que ayudan a comprender los datos.
- Deja que una calculadora o un ordenador haga la mayor parte de los cálculos y los gráficos, de manera que puedas concentrarte en qué haces y por qué lo haces.
- Céntrate en las grandes ideas de la estadística, no te centres sólo en las reglas y las fórmulas.

De todas formas, el principio básico del aprendizaje es la constancia. Las principales ideas de la estadística, al igual que las principales ideas de cualquier campo de estudio importante, requieren mucho tiempo para ser descubiertas y algo más de tiempo para dominarlas. Los resultados compensarán el esfuerzo.

PARTE I
COMPRENSIÓN DE LOS DATOS

Para comprender la información contenida en un conjunto de datos hay que empezar por organizar y presentar los datos de tal forma que "hablen por sí mismos". Esto es el *análisis de datos*. El paso siguiente consiste en prestar una cuidadosa atención al origen de los datos. Esto es la *obtención de datos*. El análisis y la obtención de datos son los puntos de partida de la *inferencia estadística*, cuyo objetivo consiste en extender a un colectivo más amplio las conclusiones obtenidas con los individuos concretos que describen nuestros datos. Los tres capítulos de la primera parte tratan del análisis y obtención de datos.

Los capítulos 1 y 2 reflejan la gran importancia que se da al análisis de datos en la estadística aplicada moderna. Aunque el análisis cuidadoso de los datos es imprescindible para la inferencia estadística, el análisis de datos es algo más que el prólogo de la inferencia. En realidad, hay que distinguir claramente entre los datos de que disponemos y el universo más amplio al que queremos extender nuestras conclusiones. Por ejemplo, en Estados Unidos la tasa de desempleo se determina a partir de una encuesta a 60.000 hogares, aunque el objetivo sea el de sacar conclusiones referidas a la totalidad de los 96 millones de hogares de aquel país. Este es un problema complejo, como veremos en la segunda parte de este libro.

Desde el punto de vista del análisis de datos las cosas son más simples. Nos basta con explorar y comprender los datos de que disponemos, sin preocuparnos de su origen. En los capítulos 1 y 2 estudiaremos una estrategia sistemática para examinar datos, y presentaremos los instrumentos necesarios para poner en práctica esta estrategia.

Aunque, por supuesto, a menudo queremos utilizar los datos para alcanzar conclusiones más generales, el que esto sea posible depende sobre todo de cómo se obtuvieron. Los datos raramente "caen del cielo"; son producto del esfuerzo humano, como las medias de nailon o las gafas de sol. El capítulo 3 nos enseña cómo obtener buenos datos y cómo decidir si podemos confiar en los que han obtenido los demás.

El estudio del análisis y obtención de datos te proporciona ideas y herramientas que serán de gran utilidad cuando tengas que vértelas con los números. La inferencia es algo más especializada y sutil. Exige que el libro de texto le dedique más atención, pero eso no significa que sea más importante. La estadística es la ciencia de los datos y los tres capítulos de esta primera parte tratan directamente sobre ellos.

1. ANÁLISIS DE DISTRIBUCIONES

FLORENCE NIGHTINGALE

A Florence Nightingale (1820-1910) se la conoce por ser fundadora de la profesión de enfermería, y por su importante labor como reformadora del sistema de atención sanitaria del ejército británico. Como enfermera jefe de dicho ejército durante la Guerra de Crimea, de 1854 a 1856, Florence se percató de que la falta de medidas sanitarias era la causa principal del fallecimiento de muchos soldados heridos en combate. Con las reformas que Nightingale introdujo en el hospital militar donde trabajaba, la tasa de mortalidad pasó del 42.7% al 2.2%. Cuando Nightingale volvió a Gran Bretaña inició, con considerable éxito, una feroz lucha para reformar todo el sistema de atención sanitaria.

Una de las armas que Florence Nightingale utilizó para conseguir sus propósitos fueron los datos. Florence no sólo modificó el sistema de atención sanitaria, sino que también modificó el sistema de registro de datos. Los datos de que disponía le sirvieron para respaldar sus argumentos de forma muy sólida. Nightingale fue una de las primeras personas en utilizar gráficos para representar datos de forma sencilla, de tal manera que incluso los generales y los miembros del parlamento podían entenderlos. Sus representaciones gráficas de los datos constituyen un hito en el desarrollo de la estadística como ciencia. Florence Nightingale consideró que la estadística era esencial para poder comprender cualquier fenómeno social e intentó introducirla en la educación superior.

Al empezar a estudiar estadística, queremos seguir el camino que inició Florence Nightingale. En este capítulo y en el siguiente, daremos especial importancia al análisis de datos. Como hizo Nightingale, empezaremos representando los datos gráficamente. A los gráficos les añadiremos algunos cálculos numéricos, como también hizo Nightingale al calcular tasas de mortalidad. Para Florence Nightingale los datos no eran algo abstracto ya que le permitían comprender, y hacer comprender a los demás, la forma de salvar vidas humanas. Lo mismo puede decirse en la actualidad.

1.1 Introducción

La estadística es la ciencia de los datos. Por lo tanto, empezamos nuestro estudio de la estadística adentrándonos en el arte de examinar datos. Cualquier conjunto de datos contiene información sobre un grupo de *individuos*. La información se organiza en forma de *variables*.

> INDIVIDUOS Y VARIABLES
>
> Los **individuos** son las personas, animales o cosas descritos en un conjunto de datos.
>
> Una **variable** es cualquier característica de un individuo. Las variables pueden tomar distintos valores para los distintos individuos.

Por ejemplo, los datos para el estudio de la política salarial de una empresa tienen que hacer referencia a todos los empleados. Estos son los individuos descritos por el conjunto de datos. Para cada individuo, los datos contienen los valores de variables como la edad en años, el sexo (hombre o mujer), la categoría laboral o el sueldo. En la práctica, cualquier conjunto de datos se acompaña de una información general que ayuda a comprenderlos. Cuando te encuentres con un conjunto de datos nuevo, plantéate las siguientes preguntas:

1. ¿Qué **individuos** describen los datos? ¿**Cuántos** individuos aparecen en los datos?
2. ¿Cuántas **variables** contienen los datos? ¿Cuáles son las **definiciones exactas** de dichas variables? ¿En qué *unidades* se ha registrado cada variable? El peso, por ejemplo, se puede expresar en kilogramos, en quintales o en toneladas. ¿Puede haber algún motivo para desconfiar del valor de alguna variable?
3. ¿Qué **propósito** se persigue con estos datos? ¿Queremos responder alguna pregunta concreta? ¿Queremos obtener conclusiones sobre unos individuos de los que no tenemos realmente datos?

El tercer grupo de preguntas es muy importante, tan importante que se trata en profundidad en el capítulo 3. Sin embargo, de momento nos daremos por satisfechos describiendo a los individuos y las variables de un conjunto de datos.

EJEMPLO 1.1

He aquí una pequeña parte de un conjunto de datos que describe a los Estados europeos.

Estado	Región	Población (1.000 hab.) 1993	Superficie (km^2)	PIB per cápita 1994 (dólares)	Periódicos (1.000 hab.)	Televisores (1.000 hab.)	% PIB en educación pública
⋮							
Dinamarca	UE	5.165	43.069	28.110	332	538	7.40
Eslovaquia	EE	5.314	49.035	2.230	317	474	5.70
Eslovenia	EE	1.937	20.521	7.140	160	297	6.20
España	UE	39.514	504.782	13.280	104	400	4.60
⋮							

Los *individuos* descritos son los Estados europeos. Se presentan datos de cuatro Estados. Cada fila de la tabla describe a un individuo. Cada columna contiene los valores que toma una *variable* para todos los individuos. Así es cómo se estructuran habitualmente las tablas de datos. La primera columna identifica los Estados. Damos datos de Dinamarca, Eslovaquia, Eslovenia y España. La segunda columna indica la región socio-política a la que pertenece cada Estado. En Europa se pueden distinguir las siguientes regiones socio-políticas: los países de la Unión Europea (UE), los países del Este (EE, ex bloque soviético) y otros países (OT). La tercera y la cuarta columnas son estimaciones de la ONU sobre la población de cada Estado en 1993 en miles de personas y sobre su superficie total. Fíjate en que las *unidades* de la población están en miles de personas y la superficie en kilómetros cuadrados. En España, 39.514 significa 39.514.000 personas. Esta población se estimó para 1993, por lo que es relativamente reciente. De todas formas, esta columna no incluye los últimos cambios que se hayan podido producir.

La quinta columna contiene estimaciones del Banco Mundial sobre el producto interior bruto per cápita de cada Estado para el año 1994 expresado en dólares.

Las tres variables restantes son índices educativos y culturales utilizados por la UNESCO para caracterizar los distintos países del mundo. Las variables sexta y séptima son el número de periódicos (el promedio del número de ejemplares vendidos cada día) y el número de aparatos de televisión por cada 1.000 habitantes (este índice se basa en una estimación del número de televisiores en funcionamiento). Son estimaciones para los años 1992 y 1993, respectivamente. Finalmente, la última columna contiene una estimación del gasto público en educación de cada Estado para el año 1993 expresado como un porcentaje sobre la renta per cápita. ■

Análisis exploratorio de los datos

Tal como muestra el ejemplo 1.1, una comprensión total de las variables de un conjunto de datos a menudo exige algún estudio general de las variables. Afortunadamente, cuando trabajas con datos propios sueles conocer bien las variables. Los conceptos y las ideas estadísticas te pueden ayudar a examinar tus datos para describir sus características principales. Este examen se llama ***análisis exploratorio de los datos***. Al igual que un explorador que cruza tierras desconocidas, lo primero que haremos será, simplemente, describir lo que vemos. Cada ejemplo que presentemos tendrá alguna información general que nos ayudará a entenderlo. De todas formas, nos centraremos en el examen de los datos. He aquí dos estrategias básicas que nos ayudan a organizar nuestra exploración de un conjunto de datos:

- Empieza examinando cada variable de forma independiente. Luego, pasa al estudio de las relaciones entre variables.
- Empieza con uno o varios gráficos. Luego, añade resúmenes numéricos de aspectos concretos de los datos.

Organizaremos nuestro aprendizaje de la misma manera. Este capítulo hace referencia al examen de una sola variable y el capítulo 2 examina las relaciones entre variables. En cada capítulo empezamos con gráficos y luego pasamos a los resúmenes numéricos.

1.2 Presentación de distribuciones con gráficos

Algunas variables, como el sexo o la profesión, simplemente clasifican a los individuos en categorías. Otras, en cambio, como la estatura o los ingresos anuales, toman

VARIABLES CATEGÓRICAS Y VARIABLES CUANTITATIVAS

Una **variable categórica** indica a qué grupo o a qué categoría pertenece un individuo.

Una **variable cuantitativa** toma valores numéricos, para los que tiene sentido hacer operaciones aritméticas como sumas y promedios.

La **distribución** de una variable nos dice qué valores toma una variable y con qué frecuencia.

valores numéricos con los que podemos hacer cálculos aritméticos. Tiene sentido dar un promedio de los ingresos de los trabajadores de una empresa, pero no tiene sentido dar un sexo "promedio". Podemos, sin embargo, hacer un recuento de los hombres y mujeres empleados y hacer cálculos con estos recuentos.

1.2.1 Variables categóricas

Los valores de una variable categórica son simplemente etiquetas asignadas a las categorías de la misma como, por ejemplo, "hombre" y "mujer". La distribución de una variable categórica lista las categorías y da el recuento o el porcentaje de individuos de cada categoría. Por ejemplo, he aquí la distribución del número de familias por tipos en Suecia según datos del Eurostat de 1991.

Tipos de familia	Recuento (miles)	Porcentaje
Parejas sin hijos	1.168	53.50
Parejas con hijos	830	38.02
Hombres solos con hijos	27	1.24
Mujeres solas con hijos	158	7.24

Diagramas de barras y diagramas de sectores

Para presentar datos como los que hemos visto, por ejemplo en una conferencia, puede ser que desees utilizar gráficos como los de la figura 1.1. El *diagrama de barras* de la figura 1.1(a) compara de forma rápida el tamaño de los cuatro tipos de familias. Las alturas de las cuatro barras muestran el número de individuos de cada categoría. El *diagrama de sectores* de la figura 1.1(b) nos ayuda a ver la importancia relativa de cada categoría respecto al total. Por ejemplo, se ve que la porción de "parejas sin hijos" corresponde al 53.5% del total, ya que el 53.5% de las familias son parejas sin hijos. Los diagramas de barras y los de sectores ayudan a captar de forma rápida la distribución de una variable categórica. Aunque nos facilitan la comprensión de los datos, estos diagramas no son imprescindibles. De hecho, cuando las variables categóricas se analizan de forma aislada, como por ejemplo el tipo de familia, se pueden describir fácilmente sin la ayuda de ningún gráfico. Tienes que ser capaz de interpretar los diagramas de barras y de sectores, pero nosotros no los utilizaremos. Podemos pasar directamente a los gráficos con variables cuantitativas.

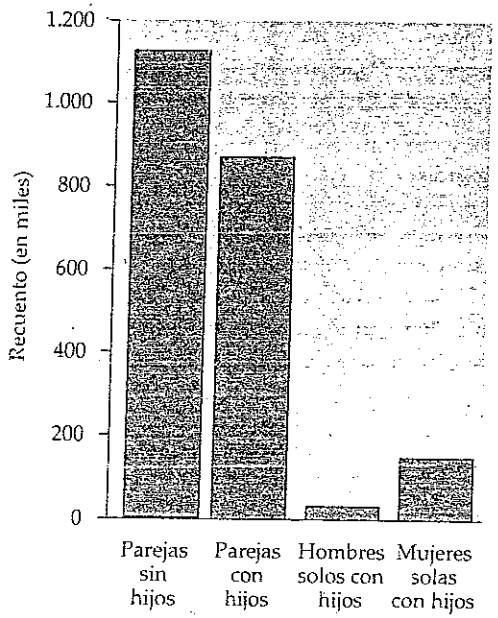

Figura 1.1(a). Diagrama de barras del número de familias por tipos en Suecia según datos del Eurostat de 1991.

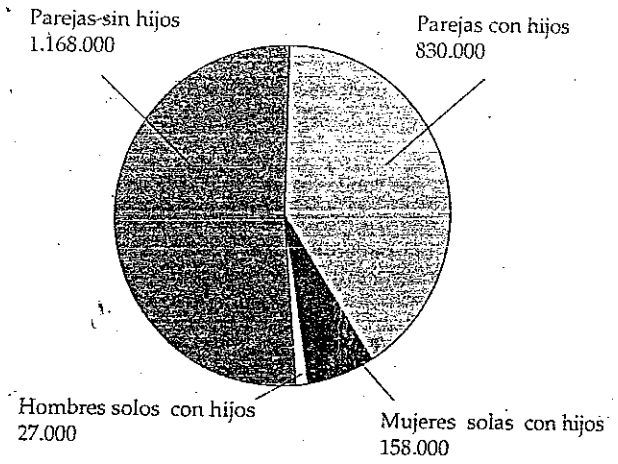

Figura 1.1(b). Diagrama de sectores de los mismos datos.

EJERCICIOS

1.1. El ejemplo 1.1 presenta datos referentes a los Estados europeos. La primera columna identifica el Estado. Cada una de las siete columnas siguientes contiene valores de otras variables. ¿Cuáles de estas variables son categóricas y cuáles cuantitativas?

1.2. Un estudio médico describe un conjunto de variables referentes a un grupo de pacientes. De las variables siguientes, ¿cuáles son categóricas y cuáles son cuantitativas?

(a) Sexo (hombre o mujer)
(b) Edad (años)
(c) Raza (asiática, blanca, negra u otras)
(d) Fumador (sí o no)
(e) Presión sanguínea sistólica (milímetros de mercurio)
(f) Nivel de calcio en la sangre (microgramos por mililitro)

1.2.2 Dibujo de histogramas

Las variables cuantitativas a menudo toman tantos valores que un gráfico de su distribución es más claro si se agrupan los valores próximos. El gráfico más común de la distribución de una variable cuantitativa es un *histograma*.

EJEMPLO 1.2

La tabla 1.1 presenta los porcentajes de residentes mayores de 65 años en cada uno de los 50 Estados de EE UU. Para dibujar un histograma de esta distribución procede de la manera siguiente:

1. Divide el recorrido de los datos (diferencia entre los valores máximo y mínimo) en clases de igual amplitud. Los datos de la tabla 1.1 van desde 4.2 hasta 18.3, por lo que escogemos como nuestras clases:

$$4{,}0 < \text{porcentaje de mayores de 65 años} \leq 5{,}0$$
$$5{,}0 < \text{porcentaje de mayores de 65 años} \leq 6{,}0$$

$$18{,}0 < \text{porcentaje de mayores de 65 años} \leq 19{,}0$$

Asegúrate de especificar las clases con precisión, de manera que cada observación se sitúe exactamente en una clase. Un Estado con un 5% de sus residentes mayores de

65 años se situará en la primera clase, pero un Estado con un 5,1% se situará en la segunda clase.

Tabla 1.1. Porcentaje de la población mayor de 65 años en cada Estado de EE UU (1991).

Estado	Porcentaje	Estado	Porcentaje
Alabama	12,9	Michigan	12,1
Alaska	4,2	Minnesota	12,5
Arizona	13,2	Misisipí	12,4
Arkansas	14,9	Misurí	14,1
California	10,5	Montana	13,4
Carolina del Norte	12,3	Nebraska	14,1
Carolina del Sur	11,4	Nevada	10,8
Colorado	10,1	New Hampshire	11,6
Connecticut	13,7	Nueva Jersey	13,4
Dakota del Norte	14,5	Nuevo México	10,9
Dakota del Sur	14,7	Nueva York	13,1
Delaware	12,2	Ohio	13,1
Florida	18,3	Oklahoma	13,5
Georgia	10,1	Oregón	13,7
Hawai	11,4	Pensilvania	15,5
Idaho	12,0	Rhode Island	15,1
Illinois	12,5	Tejas	10,1
Indiana	12,6	Tennessee	12,7
Iowa	15,4	Utah	8,8
Kansas	13,9	Vermont	11,9
Kentucky	12,7	Virginia	10,9
Luisiana	11,2	Virginia Occidental	15,1
Maine	13,4	Washington	11,8
Maryland	10,9	Wisconsin	13,3
Massachusetts	13,7	Wyoming	10,6

Fuente: *Statistical Abstract of the United States*, 1992.

2. Haz un recuento del número de observaciones de cada clase. He aquí los recuentos.

Clase	Recuento	Clase	Recuento	Clase	Recuento
4,1 a 5,0	1	9,1 a 10,0	0	14,1 a 15,0	5
5,1 a 6,0	0	10,1 a 11,0	9	15,1 a 16,0	4
6,1 a 7,0	0	11,1 a 12,0	7	16,1 a 17,0	0
7,1 a 8,0	0	12,1 a 13,0	10	17,1 a 18,0	0
8,1 a 9,0	1	13,1 a 14,0	12	18,1 a 19,0	1

3. Dibuja el histograma. Primero marca la escala de valores de la variable cuya distribución muestras en el eje de las abcisas. En este ejemplo, es el "porcentaje de residentes mayores de 65 años". La escala va de 4 a 19. Seguidamente, marca la escala de recuentos en el eje de las ordenadas. Cada barra representa a una clase. La amplitud de la barra debe cubrir todos los valores de la clase. La altura de la barra es el número de observaciones de cada clase. No dejes espacios vacíos entre barras (a no ser que alguna clase este vacía y que, por lo tanto, su barra tenga altura cero). La figura 1.2 es nuestro histograma. ■

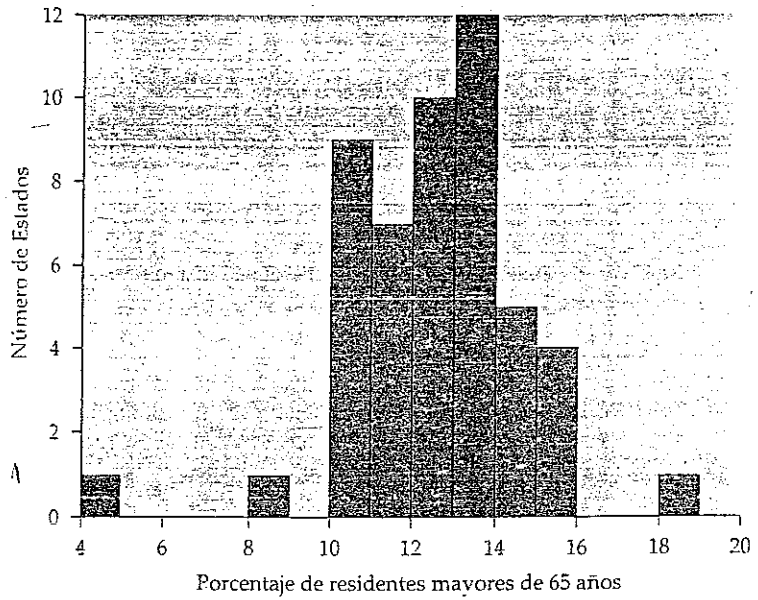

Figura 1.2. Histograma del porcentaje de residentes mayores de 65 años en los 50 Estados de EE UU. Datos de la tabla 1.1.

Las barras de un histograma deben cubrir todo el recorrido de una variable. Cuando falten algunos de los posibles valores de la variable, extiende las bases de las barras hasta llegar a medio camino de dos valores adyacentes posibles. Por ejemplo, en un histograma sobre las edades en años de los profesores de una universidad, las barras que representan las edades de 25 a 29 años y de 30 a 34 años se deben encontrar en 29,5. No hay una sola elección correcta para las clases de un histograma. Pocas clases pueden dar un gráfico con aspecto de "rascacielos" con todos los valores en unas pocas clases con barras altas. Demasiadas clases pueden dar un gráfico con aspecto "aplastado" con la mayoría de clases con una o ninguna observación. Ninguna de las elecciones anteriores dará una buena representación de la forma de la distribución. Cuando escojas las clases, tienes que utilizar tu sentido común para mostrar la forma

de una distribución. Nuestra vista responde al *área* de las barras de un histograma, por lo tanto, asegúrate de escoger clases que tengan la misma anchura. Entonces, el área está determinada por la altura y todas las clases están representadas de forma equitativa. Si utilizas un ordenador el programa estadístico escogerá las clases por defecto. La elección del ordenador en general es buena pero, si quieres, puedes cambiarla.

1.2.3 Interpretación de los histogramas

Hacer un gráfico estadístico no es un fin en sí mismo. Además, los ordenadores lo pueden hacer mucho más deprisa que nosotros. El objetivo de un gráfico es ayudarnos a comprender los datos. Después de que tu ordenador o tú mismo hagáis un gráfico, pregunta siempre: "¿qué veo?". He aquí una táctica general para analizar los gráficos:

- Identifica el aspecto general y también las desviaciones sorprendentes.

En el caso de los histogramas, el aspecto general es la forma de la distribución. Las *observaciones atípicas* son un tipo importante de desviaciones.

> OBSERVACIONES ATÍPICAS
>
> Una **observación atípica** de un gráfico de datos es la observación individual que no queda descrita por el aspecto general del gráfico.

En el histograma de la figura 1.2 hay tres Estados que destacan. Los puedes hallar en la tabla una vez que el histograma ha facilitado su detección. Florida tiene un 18.3% de residentes mayores de 65 años, mientras que Alaska tiene sólo un 4.2%. Estos Estados son claramente observaciones atípicas. También podrías considerar Utah, con un 8,8% de la población mayor de 65 años, una observación atípica, aunque dicho Estado no esté tan alejado del resto como Florida y Alaska. De hecho, el que una observación sea o no atípica no deja de ser algo subjetivo. Es mucho más fácil identificar observaciones atípicas en un histograma que en una tabla.

Una vez detectadas las observaciones atípicas, busca una explicación. Muchas observaciones atípicas se deben a errores de transcripción como, por ejemplo, escribir 4,0 en vez de 40. Otras veces, en cambio, las observaciones atípicas apuntan hacia alguna característica especial de ciertas observaciones. Su explicación a menudo exige tener información de carácter más general. Por ejemplo, a los estadounidenses no les sorprende que Florida, el lugar de retiro de muchos jubilados, tenga muchos residentes mayores de 65 años y que Alaska, que es la frontera norte del país, tenga pocos.

¿Qué se puede decir sobre el *aspecto general* de la figura 1.2?

> ASPECTO GENERAL DE UNA DISTRIBUCIÓN
>
> Para describir el aspecto general de una distribución:
>
> - Proporciona su **centro** y su **dispersión**.
> - Mira si la distribución tiene una **forma** simple que puedas describir en pocas palabras.

La siguiente sección explica con detalle cómo medir el centro y la dispersión. De momento, describe el centro hallando un valor que divida las observaciones de manera que aproximadamente una mitad tome valores mayores y la otra mitad valores menores. El centro de la figura 1.2 es aproximadamente el 13%. Es decir, cerca de un 13% de los residentes de un Estado típico son mayores de 65 años. Puedes describir la dispersión como la diferencia entre los valores máximo y mínimo. La dispersión en la figura 1.2 va del 10 al 16%, si ignoramos las observaciones atípicas. El histograma de la figura 1.2 tiene una forma irregular que no es fácil de describir. No obstante, algunas distribuciones tienen formas sencillas. He aquí ejemplos que ilustran algunas formas en las que hay que fijarse.

EJEMPLO 1.3

Observa los histogramas de las figuras 1.3 y 1.4. La figura 1.3 se deriva de un estudio sobre las tormentas acompañadas de aparato eléctrico en una determinada localidad de Colorado, EE UU. La figura muestra la distribución de la hora del día en que se produce el primer relámpago. La distribución tiene un solo pico a mediodía y va disminuyendo a ambos lados según nos alejamos de este pico. Los dos lados del histograma tienen aproximadamente la misma forma, por ello, a esta distribución la llamaremos *simétrica*.

La figura 1.4 muestra la distribución de la longitud de las palabras utilizadas en las obras de Shakespeare.[1] Esta distribución es *asimétrica hacia la derecha*. Es decir, hay muchas palabras cortas (de 3 o 4 letras) y muy pocas largas (10, 11 o 12 letras), de manera que la cola de la derecha del histograma se extiende mucho más lejos que la cola de la izquierda.

[1] C. B. Williams, *Style and Vocabulary: Numerical Studies*. Griffin, Londres, 1970.

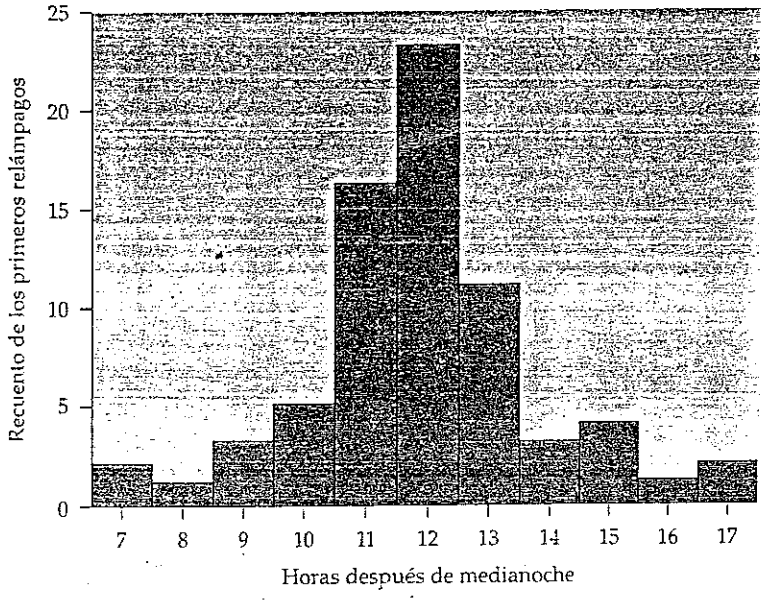

Figura 1.3. Distribución de la hora en la que se produce el primer relámpago del día en una localidad de Colorado, EE UU.

Fíjate en que la escala del eje de las ordenadas de la figura 1.4 no es un *recuento* de palabras, sino que es el *porcentaje* de todas las palabras de Shakespeare con una determinada longitud. Un histograma de porcentajes es más conveniente que un histograma de recuentos cuando tenemos muchas observaciones, o cuando queremos comparar varias distribuciones. Diferentes estilos literarios tienen distintas distribuciones de la longitud de las palabras empleadas, pero todas ellas son asimétricas hacia la derecha, ya que las palabras cortas son frecuentes y las palabras muy largas lo son menos. ■

En matemáticas, simetría significa que los dos lados de una figura, como un histograma, son imágenes especulares exactas la una de la otra. Las distribuciones de datos casi nunca son exactamente simétricas. De todas formas, en general diremos que histogramas como el de la figura 1.3 son aproximadamente simétricos.

La forma de una distribución nos da información importante sobre una variable. Algunos tipos de datos generan sistemáticamente distribuciones que son simétricas o asimétricas. Por ejemplo, los tamaños de muchos individuos distintos de una misma especie (como las longitudes de las cucarachas) tienden a ser simétricos. Los datos sobre los ingresos (de personas, empresas o Estados) son, normalmente, muy

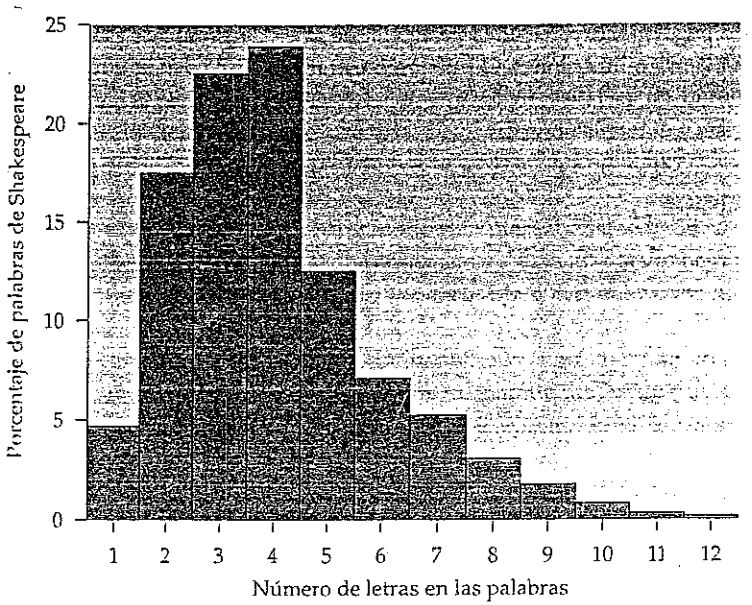

Figura 1.4. Distribución de la longitud de las palabras utilizadas en las obras de Shakespeare.

DISTRIBUCIONES SIMÉTRICAS Y DISTRIBUCIONES ASIMÉTRICAS

Una distribución es **simétrica** si los lados derecho e izquierdo del histograma son aproximadamente imágenes especulares el uno del otro.

Una distribución es **asimétrica hacia la derecha** si el lado derecho del histograma (que contiene la mitad de las observaciones mayores) se extiende mucho más lejos que el lado izquierdo (que contiene la mitad de las observaciones menores). Una distribución es **asimétrica hacia la izquierda** si el lado izquierdo del histograma se extiende mucho más allá que el lado derecho.

asimétricos hacia la derecha. Hay muchos ingresos moderados, algunos elevados y muy pocos ingresos muy elevados. Recuerda que muchos histogramas, como la figura 1.2, no pueden calificarse, de manera razonable, ni de simétricos ni de asimétricos. Algunos datos muestran otro tipo de formas. Por ejemplo, las calificaciones de un

examen pueden agruparse en la parte alta de la escala si muchos estudiantes obtuvieron buenas calificaciones. O pueden mostrar dos picos distintos si un problema difícil dividió a la clase entre los que lo resolvieron y los que no lo resolvieron. Utiliza la vista y di lo que observas.

EJERCICIOS

1.3. El rendimiento total de una acción se obtiene teniendo en cuenta su precio de venta en Bolsa y los dividendos pagados por la misma. El rendimiento total se expresa normalmente como un porcentaje sobre el precio de compra inicial. La figura 1.5 es un histograma de la distribución de los rendimientos totales de 1.528 acciones en la Bolsa de Nueva York durante un año.[2] Al igual que la figura 1.4, la figura 1.5 es un histograma de los porcentajes de cada clase y no un histograma de recuentos.

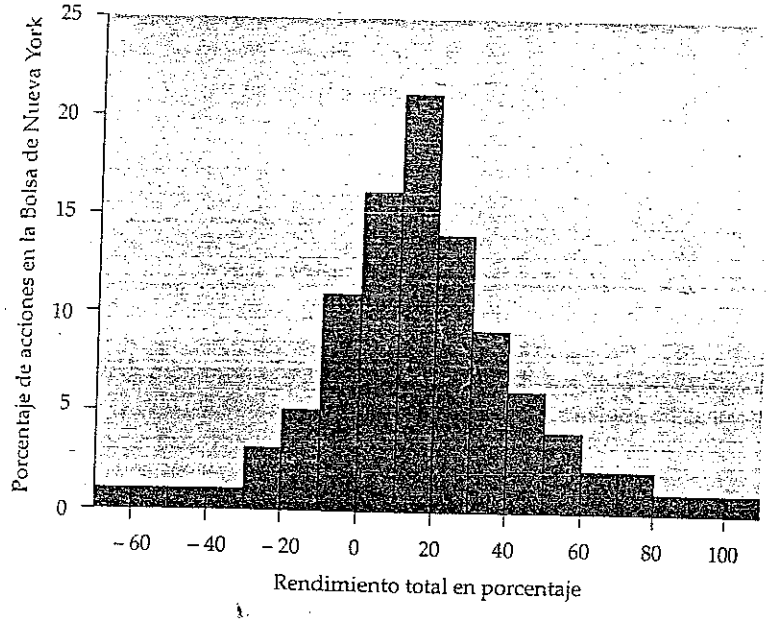

Figura 1.5. Histograma de la distribución de porcentajes de los rendimientos totales de todas las acciones en la Bolsa de Nueva York durante un año, para el ejercicio 1.3.

[2] C. H. Ford, "Diversification: how many stocks will suffice?", *American Association of Individual Investors Journal*, enero de 1990, págs. 14-16.

(a) Describe la forma de la distribución de los rendimiento totales.

(b) ¿Cuál es el centro aproximado de esta distribución? (Recuerda que, por ahora, consideramos el centro como aquel valor respecto al cual la mitad de las acciones tienen valores superiores y la otra mitad inferiores).

(c) De una manera aproximada, ¿cuáles son los rendimientos mínimo y máximo? (Estos resultados describen la dispersión de la distribución).

(d) Un rendimiento total menor que cero significa que se ha perdido dinero. ¿Qué porcentaje de las acciones ha perdido dinero?

1.4. La figura 1.6 es un histograma del número de días del mes de abril en los que se produjeron heladas en Greenwich, Inglaterra.[3] Los datos cubren un periodo de 65 años.

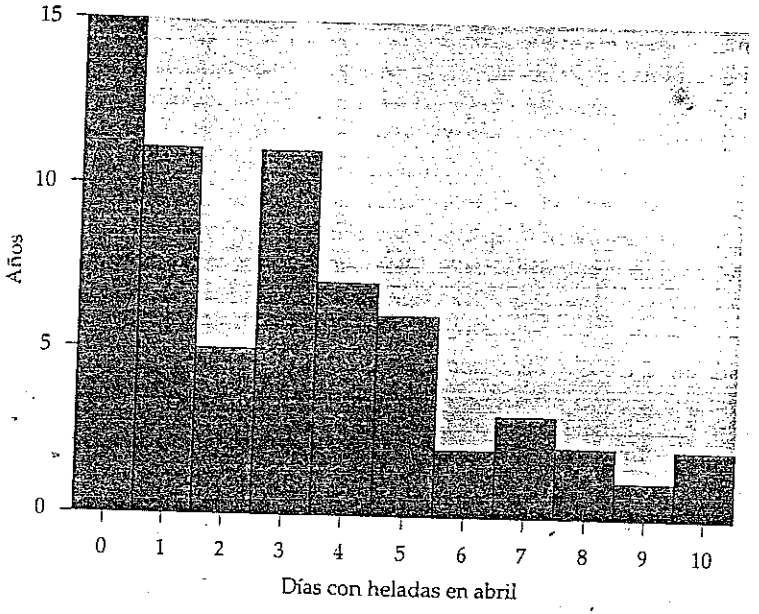

Figura 1.6. Distribución del número de días con heladas en el mes de abril en Greenwich, Inglaterra, durante un periodo de 65 años, para el ejercicio 1.4.

[3] C. E. Brooks y N. Carruthers, *Handbook of Statistical Methods in Meteorology*, H. M. Office, 1953.

(a) Describe la forma, el centro y la dispersión de esta distribución. ¿Existen observaciones atípicas?

(b) ¿Qué porcentaje de esos 65 años no tuvo ninguna helada durante el mes de abril?

1.5. ¿Cómo describirías el centro y la dispersión de la distribución de la hora del día en la que se produce el primer relámpago de la figura 1.3? ¿Y de la distribución de la longitud de las palabras en las obras de Shakespeare de la figura 1.4?

1.6. La distribución por edades de la población de una nación tiene una fuerte influencia sobre sus condiciones de vida económicas y sociales. La siguiente tabla muestra la distribución por edades de los residentes en EE UU en el año 1950 y en el 2075, en millones de personas. Los datos del año 1950 proceden del Censo de Población de ese año: Los datos del año 2075 corresponden a una predicción oficial.

(a) Como la población total del año 2075 es muy superior a la de 1950, la comparación de los porcentajes de cada grupo de edad es más clara que la comparación de los recuentos. Construye una tabla de los porcentajes de la población total en cada grupo de edad para 1950 y para 2075.

(b) Dibuja el histograma de la distribución por edades (en porcentajes) del año 1950. Luego, describe las características más importantes de esta distribución. En particular, fíjate en el porcentaje de niños respecto al total de la población.

(c) Dibuja un histograma con los datos estimados del año 2075. Utiliza las mismas escalas que has empleado en el apartado (b) para facilitar la comparación. ¿Cuáles son los cambios más importantes en la distribución por edades de la población estimada de EE UU durante el periodo de 125 años entre 1950 y 2075?

Grupo de edad	1950	2075
Menor de 10 años	29,3	34,9
De 10 a 19 años	21,8	35,7
De 20 a 29 años	24,0	36,8
De 30 a 39 años	22,8	38,1
De 40 a 49 años	19,3	37,8
De 50 a 59 años	15,5	37,5
De 60 a 69 años	11,0	34,5
De 70 a 79 años	5,5	27,2
De 80 a 89 años	1,6	18,8
De 90 a 99 años	0,1	7.7
De 100 a 109 años	–	1,7
Total	151,1	310.6

1.2.4 Diagramas de tallos

Los histogramas no son la única manera de representar gráficamente las distribuciones. Para conjuntos pequeños de datos, un ***diagrama de tallos*** es más rápido de hacer y presenta una información más detallada.

EJEMPLO 1.4

Tallos y hojas

Para hacer un diagrama de tallos, separa cada observación en un *tallo* que contenga todos los dígitos menos el del final (el situado más a la derecha) y en una *hoja*, el dígito del final. Para los porcentajes de "mayores de 65 años" de la tabla 1.1, el número entero de la observación es el tallo y el dígito del final (las décimas) es la hoja. El valor de Alabama, 12,9, tiene 12 de tallo y 9 de hoja. Los tallos pueden tener tantos dígitos como se necesiten, pero cada hoja tiene que consistir en un solo dígito.

- Sitúa los tallos de forma vertical en orden creciente de arriba abajo. Traza una línea vertical a la derecha de los tallos.
- Repasa todos los datos y sitúa cada hoja a la derecha de su tallo.
- Sitúa otra vez las hojas colocándolas esta vez en orden creciente desde cada tallo.

He aquí el diagrama de tallos con los datos de la tabla 1.1.

4	2
5	
6	
7	
8	8
9	
10	111568999
11	244689
12	01234556779
13	112344457779
14	11579
15	1145
16	
17	
18	3

Un diagrama de tallos tiene un aspecto parecido al de un histograma colocado en posición vertical. El diagrama de tallos del ejemplo 1.4 es exactamente igual al histograma de la figura 1.2, debido a que cada tallo es una clase del histograma. Los diagramas de tallos, a diferencia de los histogramas, mantienen los valores de cada observación. Interpretamos los diagramas de tallos como los histogramas, buscando caracterizar su aspecto general e identificando también las observaciones atípicas.

Redondeo

En un histograma puedes escoger las clases. Las clases (los tallos) de un diagrama de tallos te vienen dadas. Hay dos modificaciones de los diagramas de tallos que nos dan más flexibilidad a la hora de representar las distribuciones. La primera consiste en *redondear* los datos de manera que el dígito final, después del redondeo, sea adecuado como hoja. Haz esto cuando los datos tengan demasiados dígitos. Por ejemplo, datos como

3,468 2.567 2,981 1.095 . . .

tendrán demasiados tallos si tomamos los tres primeros dígitos como tallos y el dígito final como hoja. Probablemente convenga redondear estos datos como

3,5 2,6 3,0 1.1 . . .

antes de dibujar el diagrama de tallos.

División de los tallos

También puedes *dividir los tallos* para doblar su número cuando todas las hojas se sitúan sólo en unos pocos tallos. Cada tallo aparece, entonces, dos veces. Las hojas que van de 0 a 4 se sitúan en el tallo superior y las que van de 5 a 9 en el inferior. Si divides los tallos del diagrama de tallos del ejemplo 1.4, los tallos 11 y 12 serán

11	244
11	689
12	01234
12	556779

El redondeo o la división de los tallos es una decisión subjetiva, lo mismo que la elección de las clases de un histograma. El diagrama de tallos del ejemplo 1.4 no necesita ningún cambio. Los diagramas de tallos son útiles cuando se dispone de pocos datos. Cuando hay más de 100 observaciones, casi siempre es mejor decidirse por un histograma.

EJERCICIOS

1.7. La prueba SSHA (*Survey of Study Habits and Attitudes*) es una prueba psicológica que valora la motivación y la actitud de los estudiantes. Una universidad privada somete a la prueba SSHA a una muestra de 18 alumnas de primer curso. Los resultados son:

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

Dibuja un diagrama de tallos con estos datos. La forma de la distribución es irregular, lo cual es frecuente cuando se dispone sólo de un número pequeño de observaciones. ¿Has detectado observaciones atípicas? ¿Dónde se encuentra el centro de la distribución (es decir, la puntuación tal que una mitad de las puntuaciones son mayores y la otra mitad menores)? ¿Cuál es la dispersión de los datos (prescindiendo de las posibles observaciones atípicas)?

1.8. Una asociación a favor del estudio de lenguas contemporáneas hace unas pruebas que miden el nivel de comprensión del francés hablado. El intervalo de valores va de 0 a 36. He aquí las puntuaciones de 20 profesores de francés al inicio de un curso intensivo de verano.[4]

32	31	29	10	30	33	22	25	32	20
30	20	24	24	31	30	15	32	23	23

(a) Dibuja un diagrama de tallos con estas puntuaciones (utiliza la división de los tallos).

(b) Describe con palabras lo más destacable de la distribución de estos datos. ¿Existen observaciones atípicas?

[4] Datos de J. A. Wipf del Departamento de Lenguas Extranjeras y Literatura de la Purdue University en EE UU.

(c) ¿Qué puntuación situaría a un profesor en el centro de la distribución, con aproximadamente la mitad de las puntuaciones por encima y la otra mitad por debajo?

1.2.5 Gráficos temporales

Muchas variables se miden a lo largo del tiempo. Por ejemplo, podríamos medir la altura de un niño en crecimiento o el precio de una acción al final de cada mes. En estos ejemplos, nuestro interés principal son los cambios a lo largo del tiempo. Para mostrar los cambios a lo largo del tiempo se construye un *gráfico temporal*.

> **GRÁFICO TEMPORAL**
>
> Un gráfico temporal de una variable representa cada observación en relación al momento en que se midió. Se recomienda situar siempre la escala temporal en el eje de las abcisas y la variable que nos interesa en el eje de las ordenadas. Si no hay demasiados puntos, la unión de los puntos contíguos mediante segmentos facilita la visualización de la evolución de la variable a lo largo del tiempo.

EJEMPLO 1.5

He aquí datos sobre la tasa de mortalidad por cáncer en EE UU (expresada como el número de muertos por cada 100.000 personas) durante un periodo de 50 años que va desde 1940 hasta 1990.

Año	1940	1945	1950	1955	1960	1965	1970	1975	1980	1985	1990
Muertos	120,3	134.0	139.8	146,5	149,2	153,5	162.8	169,7	183,9	193,3	201,7

La figura 1.7 es un gráfico temporal de estos datos. El gráfico muestra un aumento constante de la tasa de mortalidad por cáncer durante los últimos cincuenta años. Este incremento no significa que no se haya progresado en el tratamiento del cáncer. Como el cáncer es una enfermedad que afecta básicamente a la gente mayor, la tasa de mortalidad por cáncer aumenta cuando la gente vive más años, incluso si mejora el tratamiento. ■

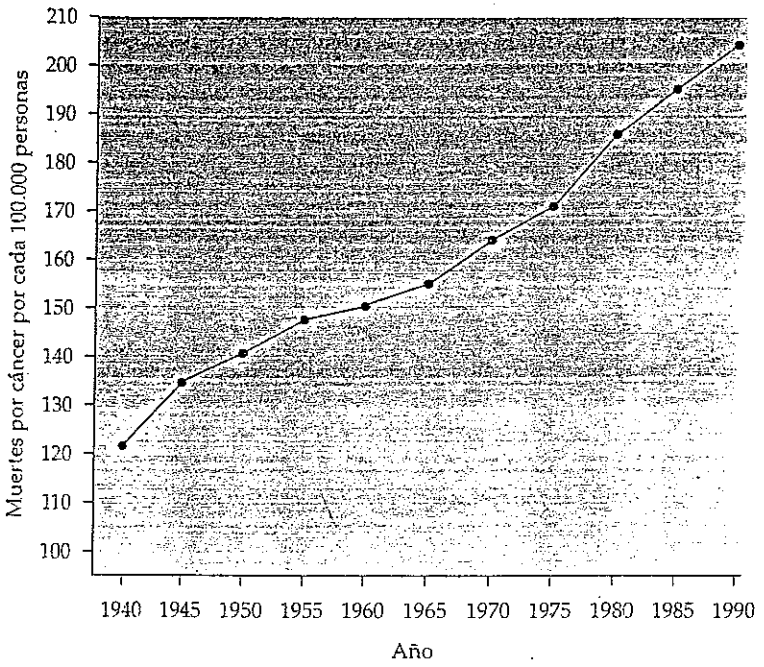

Figura 1.7. Gráfico temporal de las tasas de mortalidad por cáncer (expresado como el número de muertos por cada 100.000 personas) en EE UU, desde 1940 hasta 1990.

Tendencia

Cuando examines un gráfico temporal, fíjate una vez más en su aspecto general, en si posee una forma global bien definida y en si hay desviaciones. Una variación conjunta es una ***tendencia***; se trata de una variación, a largo plazo, creciente o decreciente. La figura 1.7 muestra una tendencia de tipo creciente en la tasa de mortalidad por cáncer, sin desviaciones notables como podrían ser disminuciones puntuales de la tasa de mortalidad.

EJERCICIOS

1.9. El Índice de Precios al Consumo (IPC) es una estadística oficial que se utiliza como referencia para tener una valoración del coste de la vida. Un IPC de 120 indica que los bienes y servicios que en el periodo de referencia costaban 100, ahora cues-

tan 120. He aquí los valores medios anuales del IPC para el periodo comprendido entre 1970 y 1989 en EE UU. El periodo de referencia va de 1982 a 1984.

Año	IPC	Año	IPC	Año	IPC	Año	IPC
1970	38,8	1975	53,8	1980	82,4	1985	107,6
1971	40,5	1976	56,9	1981	90,9	1986	109,6
1972	41,8	1977	60,6	1982	96,5	1987	113,6
1973	44,4	1978	65,2	1983	99,6	1988	118,3
1974	49,3	1979	72,6	1984	103,9	1989	124,0

(a) Dibuja un gráfico que muestre la evolución del IPC a lo largo de estos años.

(b) ¿Se observa alguna tendencia en los precios durante este periodo? ¿Hubo algún año en el que se invirtiera esta tendencia?

(c) ¿En qué épocas, durante estas décadas, hay un mayor incremento de los precios? ¿En qué periodo los aumentos de precios son menores?

RESUMEN

Un conjunto de datos contiene información sobre un número de **individuos**. Los individuos pueden ser personas, animales o cosas. Para cada individuo los datos dan valores de una o más **variables**. Una variable describe una característica de un individuo, como puede ser la altura, el sexo o el salario.

El **análisis exploratorio de datos** utiliza gráficos y resúmenes numéricos para describir las variables de un conjunto de datos y las relaciones entre ellas.

Algunas variables son **categóricas** y otras son **cuantitativas**. Una variable categórica sitúa a cada individuo en una categoría como, por ejemplo, hombre o mujer. Una variable cuantitativa tiene valores numéricos que miden alguna característica de cada individuo como, por ejemplo, la altura en centímetros o el salario anual en pesetas.

La **distribución** de una variable describe qué valores toma dicha variable y con qué frecuencia lo hace.

Para describir la distribución de una variable empieza con un gráfico. Los **histogramas** y los **diagramas de tallos** representan gráficamente las distribuciones de variables cuantitativas.

Cuando examines un gráfico o un diagrama, identifica su **aspecto general** y las **desviaciones** destacables.

El **centro**, la **dispersión** y la **forma** describen el aspecto general de una distribución. Algunas distribuciones tienen formas sencillas, como las **simétricas** y las **asimétricas**. No todas las distribuciones tienen formas sencillas, especialmente cuando hay pocas observaciones.

Las **observaciones atípicas** son observaciones que quedan fuera del aspecto general de una distribución. Busca siempre si hay observaciones atípicas e intenta explicarlas.

Cuando las observaciones de una variable corresponden a diferentes momentos del tiempo, haz un **gráfico temporal** situando la escala temporal en el eje de las abcisas y los valores de la variable en el eje de las ordenadas. Un gráfico temporal puede revelar **tendencias** u otros cambios a lo largo del tiempo.

EJERCICIOS DE LA SECCIÓN 1.2

1.10. Aquí se presenta parte de la información que una compañía posee sobre sus empleados.

Nombre	Edad	Sexo	Raza	Salario	Tipo de empleo
⋮					
Fernández, José	39	H	Blanca	153.200	Administración
Fradera, Isabel	27	M	Negra	177.400	Técnico
Fustelo, Clara	22	M	Blanca	233.750	Dirección
⋮					

(a) ¿Qué individuos describe el conjunto completo de datos?

(b) Los datos registran cinco variables a parte del nombre de cada empleado. ¿Qué variables son categóricas?

(c) ¿Qué variables son cuantitativas? Basándote en los datos de la tabla, ¿cuáles crees que son las unidades de medida de cada una de las variables cuantitativas?

1.11. El histograma de la figura 1.8 muestra el número de huracanes que alcanzaron la costa este de EE UU por año, durante un periodo de 70 años.[5] Describe brevemente la forma de esta distribución. ¿Dónde se encuentra el centro de la distribución? (Por ahora, toma el centro como el valor que tiene aproximadamente el mismo número de observaciones a cada uno de sus lados).

[5] H. C. S. Thom, 1966, *Some Methods of Climatological Analysis*, World Meteorological Organization, Ginebra, Suiza.

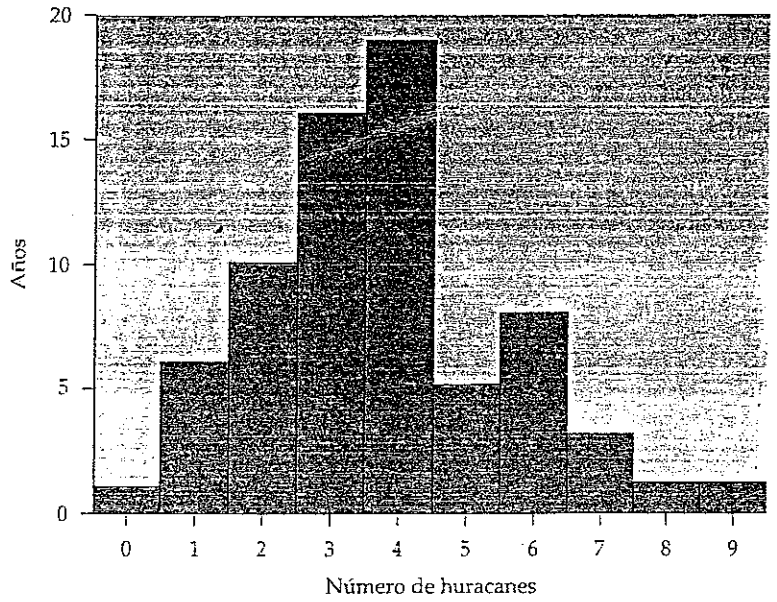

Figura 1.8. Distribución del número de huracanes que alcanzaron la costa este de EE UU por año durante un periodo de 70 años. Para el ejercicio 1.11.

1.12. La figura 1.9 presenta la distribución de los promedios de bateos de 167 jugadores de la liga americana de béisbol que batearon al menos 200 veces durante la liga de 1980 (el promedio de George Brett de 0,390 es una observación atípica; es el promedio más elevado desde el récord de 0,406 que alcanzó Ted Williams en 1941).

(a) La forma de la distribución (ignorando la observación atípica) ¿es aproximadamente simétrica, claramente asimétrica o ninguna de las dos cosas?

(b) ¿Cuál es el promedio aproximado de bateos de un jugador típico de la liga americana? ¿Cuáles son los promedios máximo y mínimo si se prescinde de la media de George Brett?

1.13. Los siguientes datos muestran el número de carreras de béisbol anuales que Babe Ruth consiguió durante 15 años (desde 1920 hasta 1934) con los New York Yankees.[6]

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

[6] Datos obtenidos de *The Baseball Encyclopedia*, 3ª ed., Macmillan, Nueva York, 1976.

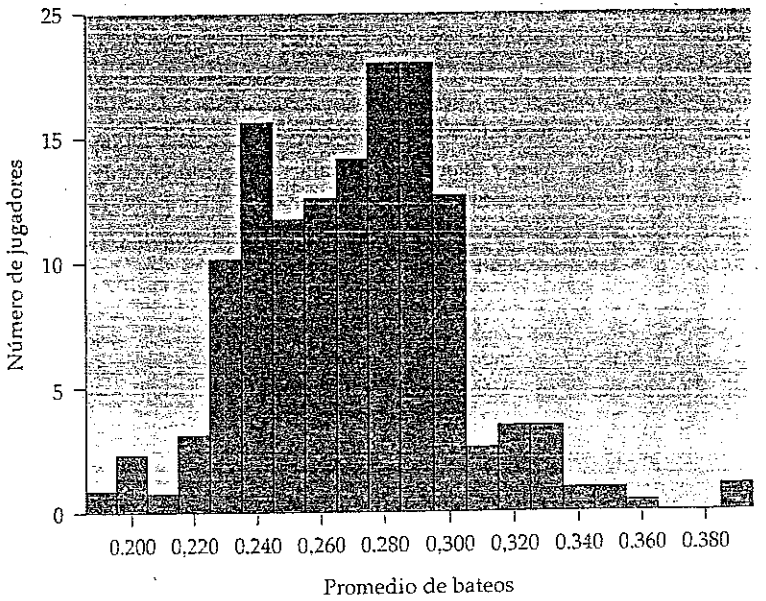

Figura 1.9. Distribución de los promedios de bateos de los 167 jugadores de la liga americana de béisbol en 1980, para el ejercicio 1.12.

Dibuja un diagrama de tallos correspondiente a estos datos. La distribución ¿es aproximadamente simétrica, claramente asimétrica o ninguna de las dos cosas? ¿Cuántas carreras consiguió Ruth en un año típico? ¿Es su famoso récord de 60 carreras en 1927 una observación atípica?

1.14 (Optativo). El récord de carreras en un solo año de Babe Ruth fue superado en 1961 por Roger Maris con 61 carreras. He aquí las carreras por año de Maris durante sus 10 años en la liga americana.

13 23 26 16 33 61 28 39 14 8

Un diagrama de tallos doble nos ayuda a comparar dos distribuciones. Parte de tu histograma del ejercicio 1.13. Traza una línea vertical a la izquierda de los tallos para dibujar un diagrama similar al que ya tienes a la derecha. Añade los datos de Maris como hojas en los mismos tallos, pero yendo hacia la izquierda en vez de hacia la derecha (asegúrate de que ordenas las hojas de cada tallo en orden creciente a partir del tallo).

¿Se puede considerar el récord de 61 carreras de Maris como una observación atípica? ¿El diagrama de tallos doble muestra que los resultados de Ruth son mejores que los de Maris?

1.15. A veces se puede obtener información útil sobre una variable representando sus valores tanto con histogramas y diagramas de tallos como con gráficos temporales. Los datos que se presentan a continuación corresponden a las *mayores* precipitaciones (en pulgadas) registradas en un día en South Bend, Indiana, de cada año de un periodo de 30. Los datos se presentan en orden cronológico.

1,88	2,23	2,58	2,07	2,94	2,29	3,14	2,15	1,95	2,51
2,86	1,48	1,12	2,76	3,10	2,05	2,23	1,70	1,57	2,81
1,24	3,29	1,87	1,50	2,99	3,48	2,12	4,69	2,29	2,12

(a) Dibuja un diagrama de tallos con estos datos. Describe la forma de la distribución. ¿Existen observaciones atípicas? (Si lo crees conveniente redondea los datos o bien dibuja el diagrama utilizando la división de tallos).

(b) Dibuja un gráfico temporal con estos datos. ¿Puede decirse que ha habido un cambio en las precipitaciones máximas registradas en South Bend?

1.16. Algunas veces un histograma o un diagrama de tallos y un gráfico temporal proporcionan información útil sobre un conjunto de datos. Los siguientes datos corresponden a mediciones de la tensión en una rejilla metálica situada detrás de la pantalla de un ordenador durante su ensamblado. Para tener una buena imagen es necesario que la tensión no sea ni demasiado alta ni demasiado baja. Por este motivo, durante la producción el fabricante controla la tensión cada hora. He aquí los promedios de estas mediciones durante 20 horas consecutivas de producción, en orden consecutivo de izquierda a derecha.

269,5	297,0	269,6	283,3	304,8	280,4	283,5	257,4	317,5	327,4
264,7	307,7	310,0	343,3	328,1	342,6	338,8	340,1	374,6	336,1

(a) Dibuja un diagrama de tallos con estos datos.

(b) Describe la forma de la distribución. La distribución presenta un variación considerable de la tensión. De todas formas, todos los valores quedan dentro de los valores correctos excepto en un caso.

(c) Representa un gráfico temporal con estos datos (marca en el eje de las abcisas las horas desde 1 hasta 20).

(d) Describe la tendencia que se observa en el gráfico temporal y explica por qué el fabricante debe iniciar una investigación.

Tabla 1.2. Datos sobre los Estados europeos.

Estado	Región	Población (1.000 hab.) 1993	Superficie (km²)	PIB per cápita 1994 (dólares)	Periódicos (por 1.000 hab.)	Televisores (por 1.000 hab.)	% PIB en educación pública
Albania	EE	3.389	28.748	360	49	89	–
Alemania	UE	80.857	356.755	25.580	323	559	–
Andorra	OT	61	453	15.000	67	367	–
Austria	UE	7.863	83.849	24.950	398	479	5,80
Bélgica	UE	10.046	30.513	22.920	310	453	5,10
Bielorrusia	EE	10.188	207.595	2.160	186	272	5,30
Bosnia-Herzegovina	EE	3.707	51.129	700	131	–	–
Bulgaria	EE	8.870	110.912	1.160	164	260	5,80
Croacia	EE	4.511	56.538	2.530	532	338	–
Dinamarca	UE	5.165	43.069	28.110	332	538	7,40
Eslovaquia	EE	5.314	49.035	2.230	317	474	5,70
Eslovenia	EE	1.937	20.521	7.140	160	297	6,20
España	UE	39.514	504.782	13.280	104	400	4,60
Estonia	EE	1.553	45.100	2.820	–	361	5,90
Finlandia	UE	5.058	337.009	18.850	512	504	7,20
Francia	UE	57.508	547.026	23.470	205	412	5,80
Grecia	UE	10.377	131.994	7.480	155	202	3,10
Holanda	UE	15.285	40.844	21.970	383	491	5,90
Hungría	EE	10.210	93.030	3.840	282	427	6,70
Irlanda	UE	3.524	70.283	13.630	186	301	6,20
Islandia	OT	263	103.000	24.590	519	335	5,60
Italia	UE	57.127	301.225	19.270	106	429	5,40
Letonia	EE	2.611	64.500	2.290	98	460	6,70
Liechtenstein	OT	30	157	35.000	653	337	–
Lituania	EE	3.712	65.200	1.350	225	383	4,40
Luxemburgo	UE	395	2.586	39.850	372	261	–
Macedonia	EE	2.119	25.713	790	27	165	5,00
Malta	OT	361	316	8.000	150	745	4,60
Moldavia	EE	4.408	33.700	870	47	–	6,50
Mónaco	OT	31	2	25.000	258	739	–
Noruega	OT	4.299	324.219	26.480	607	427	8,40
Polonia	EE	38.303	312.677	2.470	159	298	5,50
Portugal	UE	9.838	92.082	9.370	47	190	5,00
Reino Unido	UE	57.924	244.046	18.410	383	435	5,20
República Checa	EE	10.296	78.864	3.210	583	476	5,80
Rep. Fed. Yugoslavia	EE	10.623	87.968	1.000	52	179	–
Rumania	EE	23.023	237.500	1.230	324	200	3,60
Rusia	EE	147.760	17.075.400	2.650	387	372	4,40
San Marino	OT	24	61	20.000	–	352	–
Suecia	UE	8.694	449.964	23.630	511	470	8,30
Suiza	OT	7.056	41.288	23.630	377	400	5,20
Ucrania	EE	51.551	603.700	1.570	118	339	6,10

EE = Estado del Este.
UE = Unión Europea.
OT = Otros Estados.

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

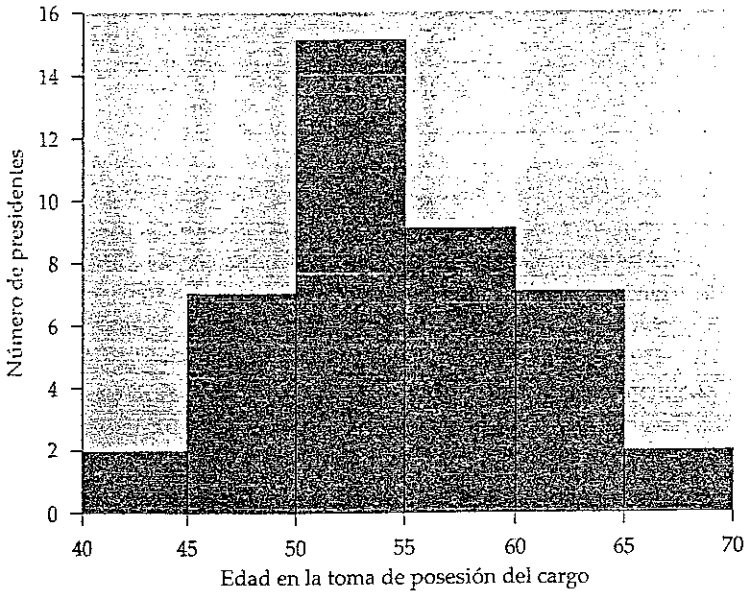

Figura 1.10. Distribución de las edades de los presidentes de EE UU en el momento de tomar posesión de su cargo. Datos de la tabla 1.3.

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. $\blacksquare$

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.

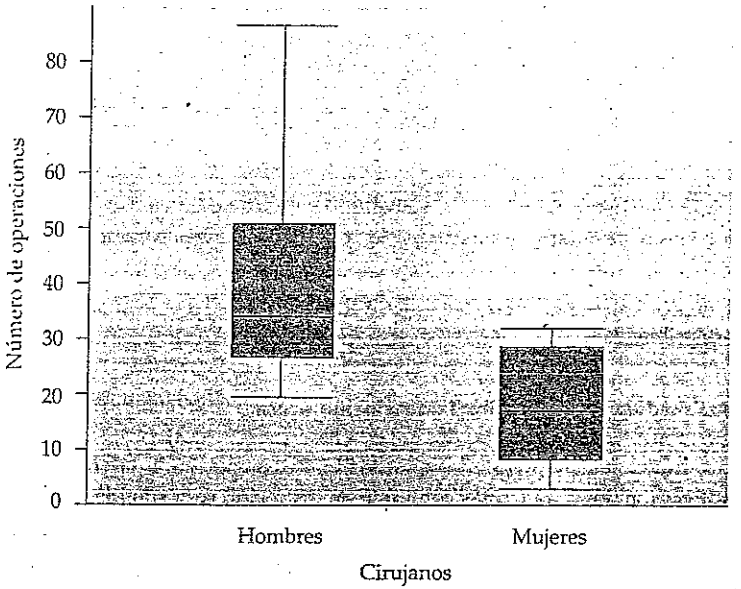

Figura 1.11. Diagramas de caja en un mismo gráfico para comparar el número de histerectomías realizadas por los cirujanos y las cirujanas.

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

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

o de una forma más simple,

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

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

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

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 $\bar{x}$ y la s de estos hombres.

En primer lugar, halla la media:

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

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

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:

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

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.

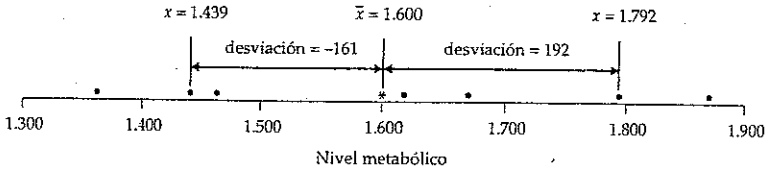

Figura 1.12. Niveles metabólicos de siete hombres, con la media (*) y las desviaciones de dos observaciones respecto a la media.

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

distribución. La mediana describe el centro, y los cuartiles y las observaciones extremas la dispersión.

Los **diagramas de caja** basados en los cinco números resumen son útiles para comparar varias distribuciones. 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 valor de la mediana se indica en el interior de la caja. Los extremos de los segmentos exteriores muestran la dispersión total de los datos.

La **varianza** s^2 y especialmente su raíz cuadrada, la **desviación típica** s, son medidas comunes de la dispersión de una distribución respecto a su media. La desviación típica s es cero cuando no hay dispersión y crece a medida que ésta aumenta.

La media y la desviación típica están muy influenciadas por las observaciones atípicas y por la asimetría de una distribución. Son buenas descripciones de las distribuciones simétricas, y son especialmente útiles en el caso de las distribuciones normales que veremos en la siguiente sección.

La mediana y los cuartiles no se ven afectados por las observaciones atípicas. Los cuartiles y los valores extremos de una distribución describen las dos colas de una distribución de forma independiente.

Los cinco números resumen son el mejor resumen numérico de las distribuciones asimétricas.

EJERCICIOS DE LA SECCIÓN 1.3

1.32. He aquí los porcentajes de votos que obtuvo cada uno de los candidatos a la Presidencia de EE UU que resultaron ganadores desde 1948 hasta 1992.

Año	1948	1952	1956	1960	1964	1968	1972	1976	1980	1984	1988	1992
Porcentajes	49.6	55.1	57.4	49.7	61.1	43.4	60.7	50.1	50.7	58.8	53.9	43.2

(a) Dibuja un diagrama de tallos correspondiente a estos porcentajes (redondea las cifras y utiliza un diagrama de tallos divididos).

(b) ¿Cuál es la mediana del porcentaje de los votos obtenidos por los candidatos que tuvieron éxito en las elecciones presidenciales? (Trabaja con los datos sin redondear).

(c) Consideraremos que fueron elecciones con victorias aplastantes aquellas en las que los porcentajes de votos se sitúan a partir del tercer cuartil. Halla el tercer cuartil. ¿En qué años se obtuvieron victorias aplastantes?

1.33. Hay gente que siempre está pendiente del número de calorías que ingiere con los alimentos. En la revista estadounidense *Consumer Reports* apareció un artículo donde se analizaban los contenidos en calorías de 20 marcas distintas de salchichas elaboradas con carne de ternera, de 17 marcas de salchichas hechas con carne de cordero, y de 17 marcas de salchichas hechas con carne de ave de corral.[7] He aquí los resultados de los análisis de los datos correspondientes a las salchichas hechas con carne de ternera:

```
Mean = 156.8    Standard deviation = 22.64    Min = 111    Max = 190
         N = 20    Median = 152.5    Quartiles = 140, 178.5
```

las salchichas hechas con carne de cordero:

```
Mean = 158.7    Standard deviation = 25.24    Min = 107    Max = 195
          N = 17    Median = 153    Quartiles = 139, 179
```

y las salchichas hechas con carne de ave de corral:

```
Mean = 122.5    Standard deviation = 24.48    Min = 87    Max = 170
          N = 17    Median = 129    Quartiles = 102, 143
```

Utiliza esta información para dibujar, en un mismo gráfico, tres diagramas de caja con los recuentos de calorías de los tres tipos de salchichas. Describe brevemente las diferencias que observes en las tres distribuciones. Comer salchichas hechas con carne de ave de corral, ¿significa ingerir menos calorías que comer salchichas hechas con carne de ternera o de cordero?

1.34. Queremos comparar el PIB de los Estados de la Unión Europea con el PIB de los Estados del Este (que formaban parte del ex bloque soviético). Como la población de los distintos Estados es muy variable, es mejor comparar el PIB per cápita en lugar de comparar directamente el PIB de cada Estado para conocer el nivel de bienestar de los ciudadanos. La tabla 1.2 da dichos valores.

(a) Haz una lista (con los valores ordenados) de los datos del PIB per cápita de los Estados de la Unión Europea y otra lista con los datos del PIB per cápita de los Estados del Este. Estas dos listas son los dos conjuntos de datos que queremos comparar.

(b) Dibuja los gráficos y calcula resúmenes numéricos para comparar las dos distribuciones. Describe brevemente lo que se observa.

[7] Este artículo apareció en el número de junio de la revista *Consumer Reports* de la pág. 366 a la 367. Un estudio más reciente sobre el mismo tema apareció en el número de julio de 1993 de la pág. 415 a la 419. Este último trabajo contiene pocos datos sobre la carne de ave de corral. Los contenidos en calorías se obtuvieron a partir de la información contenida en la etiqueta. Por tanto, los valores redondeados que se aportan son algo sospechosos.

1.35. En 1798, el científico inglés Henry Cavendish determinó la densidad de la Tierra con mucha precisión. Cuando se hacen mediciones complicadas, es aconsejable repetir la operación varias veces y trabajar con la media de todas ellas. Cavendish repitió su medición 29 veces. He aquí los resultados que obtuvo (en estos datos la densidad de la Tierra se expresa como un múltiplo de la densidad del agua).[8]

5,50	5,61	4.88	5,07	5,26	5,55	5.36	5,29	5.58	5,65
5,57	5,53	5.62	5.29	5,44	5,34	5.79	5,10	5,27	5,39
5,42	5.47	5.63	5.34	5,46	5,30	5.75	5,68	5.85	

Representa gráficamente los datos de la manera que consideres más conveniente. La forma de la distribución, ¿permite utilizar $\bar{x}$ y s para describirla? Halla $\bar{x}$ y s. Teniendo en cuenta todo lo que acabas de hacer, ¿cuál es tu estimación de la densidad de la Tierra a partir de estas mediciones?

1.36. La tabla 1.1 da el porcentaje de personas mayores de 65 años que viven en cada uno de los Estados de EE UU. La figura 1.2 es un histograma de estos datos. Para estos datos, ¿qué prefieres, los cinco números resumen o la media y la desviación típica como descripción numérica breve? ¿Por qué? Haz los cálculos que consideres más apropiados.

1.37. En 1993, los New York Mets obtuvieron el peor récord de la historia de la liga americana de béisbol. Se les pagó bien y, en cambio, jugaron muy mal. A continuación se presentan los salarios de los Mets en miles de dólares (6.200 representa un salario de 6.200.000 dólares).[9]

6.200	5.917	4.000	3.375	3.000	2.312	2.300	2.150	2.100
1.500	1.012	850	650	635	500	475	220	205
195	195	158	145	109	109	109	109	109

Describe la distribución de salarios gráficamente y con un resumen numérico apropiado. A continuación, da una breve descripción de las principales características de la distribución.

1.38. Un estudio sobre las indemnizaciones dictadas por un tribunal civil (daños personales, responsabilidad civil, etc.) de Chicago puso de manifiesto que la indemnización mediana era de 8.000 dólares. Sin embargo, la indemnización media era de 69.000 dólares. Intenta explicar las diferencias entre estas dos medidas de centro de la distribución.

[8] S. M. Stigle, "Do robust estimators work with real data?", *Annals of Statistics*, 5, 1977, págs. 1.055-1.078.

[9] Los salarios de los Mets aparecieron en el *New York Times* del 11 de abril de 1993.

1.39. En cada una de las siguientes situaciones, ¿qué medida de centro utilizarías, la media o la mediana?

(a) El Ayuntamiento de Medianilla está estudiando la posibilidad de aplicar un impuesto sobre la renta a los habitantes del pueblo. Para ello, quiere conocer el promedio de la renta de estos ciudadanos y tomarlo como referencia para el cálculo de la base impositiva.

(b) En un estudio sobre el nivel de vida de los hogares de Medianilla, un sociólogo estima la renta del hogar típico.

1.40. Supongamos que quieres determinar el promedio de la velocidad de los automóviles que circulan por una autopista en la que estás circulando. Supongamos que ajustas la velocidad de tu automóvil de manera que el número de automóviles que avanzas sea igual al número de automóviles que te avanzan. ¿Qué has determinado, la media o la mediana de la velocidad de los automóviles de la autopista?

1.41. Vamos a hacer un ejercicio sobre la desviación típica. Debes escoger cuatro números entre el 0 y el 10 (se pueden escoger números repetidos) de manera que:

(a) La desviación típica de estos números sea la más pequeña posible.

(b) La desviación típica de estos números sea la mayor posible.

(c) ¿Hay más de una posibilidad en (a) y (b)?

1.4 Distribuciones normales

Ahora disponemos de un conjunto de herramientas gráficas y numéricas para describir las distribuciones. Es más, disponemos de una estrategia clara para explorar los datos de una variable cuantitativa:

- Empezamos con un gráfico, habitualmente un diagrama de tallos o un histograma.
- Identificamos su aspecto general y las desviaciones sorprendentes, así como las observaciones atípicas.
- Escogemos un resumen numérico para describir de forma breve el centro y la dispersión de la distribución.

He aquí un elemento más que añadir a esta estrategia.