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 sobre la estadística adentrándonos en el arte de examinarlos. 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 los objetos descritos por un conjunto de datos. Los individuos pueden ser personas, pero también pueden ser animales o cosas.

Una **variable** es cualquier característica de un individuo. Una variable puede tomar distintos valores para distintos individuos.

Una base de datos sobre estudiantes universitarios, por ejemplo, contiene datos sobre cada uno de los estudiantes matriculados. Los estudiantes son los individuos descritos por el conjunto de datos. Para cada individuo, los datos contienen valores de variables como la fecha de nacimiento, el sexo (hombre o mujer), la carrera escogida o sus notas. En la práctica, cualquier conjunto de datos se acompaña de información general que ayuda a comprenderlos. Cuando planees un estudio estadístico o cuando te encuentres ante un conjunto de datos nuevo, plantéate las siguientes preguntas:

1. **¿Quién?** ¿Qué **individuos** describen los datos? ¿**Cuántos** individuos aparecen en los datos?
2. **¿Qué?** ¿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.
3. **¿Por qué?** ¿**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?

Algunas variables, como el sexo o la profesión, simplemente clasifican a los sujetos en categorías. Otras, en cambio, como la estatura o los ingresos anuales, toman valores numéricos con los que podemos hacer cálculos aritméticos. Tiene

sentido hallar la media de ingresos de los trabajadores de una empresa, pero no tiene sentido calcular un sexo "medio". Podemos, sin embargo, hacer un recuento de los hombres y mujeres empleado, y hacer cálculos con estos recuentos.

VARIABLES CATEGÓRICAS Y VARIABLES NUMÉRICAS

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

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

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

EJEMPLO 1.1. Datos sobre una empresa

He aquí una pequeña parte de un conjunto de datos sobre los empleados de una empresa:

Nombre	Edad	Sexo	Raza	Salario	Tipo de trabajo
Fleetwood, Delores	39	Mujer	Blanca	62.100	Directivo
Perez, Juan	27	Hombre	Blanca	47.350	Técnico
Wang, Lin	22	Mujer	Asiática	18.250	Administrativo
Johnson, LaVerne	48	Hombre	Negra	77.600	Directivo

Los *individuos* descritos son los empleados. Cada fila describe a un individuo. A menudo, a cada fila de datos se le llama un **caso**. Cada columna contiene los valores de una *variable* para todos los individuos. Además del nombre de cada persona, hay 5 variables. Sexo, raza y tipo de trabajo son variables categóricas. Edad y salario son variables numéricas. Observa que la edad se expresa en años y el salario en euros.

Casos

Muchas tablas de datos siguen este formato —cada fila es un individuo y cada columna es una variable—. Estos datos se presentan en una **hoja de cálculo** que contiene filas y columnas preparadas para su utilización. Las hojas de cálculo se utilizan frecuentemente para entrar y transmitir datos. ■

Hoja de cálculo

APLICA TUS CONOCIMIENTOS

1.1. He aquí un pequeño conjunto de datos sobre el consumo (en litros a los 100 kilómetros) de vehículos de 1998:

Marca y modelo	Tipo de vehículo	Tipo de cambio	Número de cilindros	Consumo en ciudad	Consumo en carretera
⋮					
BMW 318I	Pequeño	Automático	4	10,8	7,6
BMW 318I	Pequeño	Manual	4	10,3	7,4
Buick Century	Medio	Automático	6	11,8	8,2
Chevrolet Blazer	Todoterreno	Automático	6	14,8	11,8
⋮					

(a) ¿Qué individuos describe este conjunto de datos?

(b) Para cada individuo, ¿qué variables se dan? ¿Cuáles de estas variables son categóricas y cuáles numéricas?

1.2. Los datos sobre un estudio médico contienen valores de muchas variables para cada uno de los sujetos del estudio. De las siguientes variables, ¿cuáles son categóricas y cuáles son numéricas?

(a) Género (hombre o mujer).

(b) Edad (años).

(c) Raza (asiática, negra, blanca u otras).

(d) Fumador (sí, no).

(e) Presión sanguínea (en milímetros de mercurio).

(f) Concentración de calcio en la sangre (en microgramos por litro).

1.2 Gráficos de distribuciones

Análisis exploratorio de datos

Las herramientas y las ideas estadísticas nos ayudan a examinar datos para describir sus características principales. Este examen se llama **análisis exploratorio de datos**. Al igual que un explorador que cruza tierras desconocidas, lo primero que haremos será, simplemente, describir lo que vemos. Tenemos dos estrategias básicas que nos ayudan a organizar nuestra exploración de un conjunto de datos:

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

Seguiremos estos principios para organizar nuestro aprendizaje. Este capítulo hace referencia al examen de una sola variable. En el segundo capítulo estudiaremos relaciones entre varias variables. En cada capítulo empezamos con gráficos y luego pasamos a resúmenes numéricos para tener una descripción más completa.

1.2.1 Variables categóricas: diagramas de barras y diagramas de sectores

Los valores de una variable categórica son 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.

Tipo 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

Diagrama de sectores

Los gráficos de la figura 1.1 describen estos datos. El **diagrama de barras** de la figura 1.1(a) compara de forma rápida la frecuencia de los cuatro tipos de familias. La altura de las cuatro barras muestra el número de individuos de cada categoría. El **diagrama de sectores** de la figura 1.1(b) nos ayuda a visualizar 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. Para dibujar un diagrama de sectores, tienes que incluir todas las categorías que constituyen el total. Los diagramas de barras son más flexibles. Por ejemplo, puedes utilizar uno para comparar el número de estudiantes de tu universidad que se gradúan en Biología, Empresariales o Políticas. No se puede hacer esta comparación con un diagrama de sectores ya que no todos los estudiantes de la universidad pertenecen a una de estas categorías.

Los diagramas de barras, así como los de sectores, ayudan a captar de forma rápida la distribución de una variable categórica. Pero 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 pasa por ejemplo con el tipo de familia, se pueden describir fácilmente sin la ayuda de ningún gráfico. En la siguiente sección estudiaremos las variables cuantitativas, para las cuales los gráficos son herramientas esenciales.

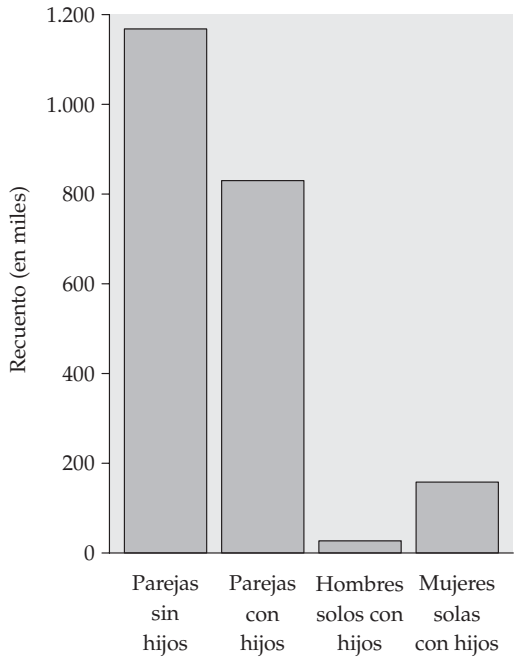

Figura 1.1(a). Diagrama de barras del número de familias por tipos en Suecia.

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

APLICA TUS CONOCIMIENTOS

1.3. Doctoras. Los datos sobre el porcentaje de mujeres que se doctoraron en distintas disciplinas en EE UU durante 1994 (según el *1997 Statistical Abstract of the United States)* son los siguientes:

Informática	15,4%	Biología	40,7%
Pedagogía	60,8%	Física	21,7%
Ingeniería	11,1%	Psicología	62,2%

(a) Presenta estos datos en forma de diagrama de barras.

(b) ¿Sería también correcto utilizar un diagrama de sectores para mostrar estos datos? Justifica tu respuesta.

1.4. Defunciones en los hospitales españoles. Según datos del Instituto Nacional de Estadística (INE) las causas de muerte más significativas en los hospitales españoles en 1996 fueron

Trastornos del aparato circulatorio	133.499
Tumores	89.204
Trastornos del aparato respiratorio	34.718
Trastornos del aparato digestivo	18.861
Trastornos del sistema inmunológico (incluye sida)	5.504
Causas externas de traumatismos y envenenamientos (incluye accidentes de tráfico)	16.324

(a) Halla el porcentaje de cada una de las causas de defunción y exprésalo con valores enteros. ¿Qué porcentaje de defunciones se debió a tumores?

(b) Dibuja un diagrama de barras de la distribución de las causas de muerte en los hospitales españoles. Identifica bien cada barra.

(c) ¿También sería correcto utilizar un diagrama de sectores para representar los datos? Justifica tu respuesta.

1.2.2 Variables cuantitativas: histogramas

Cuando las variables cuantitativas toman muchos valores, el gráfico de la distribución es más claro si se agrupan los valores próximos. El gráfico más común para describir la distribución de una variable cuantitativa es un **histograma**.

Histogramas

Tabla 1.1. Porcentaje de población mayor de 65 años en cada Estado de EE UU (1996).

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

Fuente: Statistical Abstract of the United States, 1997.

EJEMPLO 1.2. Cómo dibujar un histograma

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:

Paso 1. Divide el recorrido de los datos en clases de igual amplitud. Los datos de la tabla 1.1 van desde 5,2 hasta 18,5, por lo que escogeremos como nuestras clases:

$$\begin{aligned} 5{,}0 < &\quad \text{porcentaje de mayores de 65} \quad \leq 6{,}0 \\ 6{,}0 < &\quad \text{porcentaje de mayores de 65} \quad \leq 7{,}0 \\ &\qquad\qquad \vdots \\ 18{,}0 < &\quad \text{porcentaje de mayores de 65} \quad \leq 19{,}0 \end{aligned}$$

Asegúrate de especificar las clases con precisión, de manera que cada observación se sitúe en una sola clase. Un Estado con un 6,0% de sus residentes mayores de 65 años se situará en la primera clase, pero un Estado con un 6,1% se situará en la segunda clase.

Paso 2. Haz un recuento del número de observaciones de cada clase. En nuestro ejemplo serían

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

Paso 3. Dibuja el histograma. En el eje de las abscisas representaremos primero la escala de los valores de la variable. En este ejemplo, es el "porcentaje de residentes de cada Estado de 65 o más años". La escala va de 5 hasta 19, ya que ésta es la amplitud de valores de las clases escogidas. El eje de las ordenadas expresa la escala de recuentos. Cada barra representa 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. ■

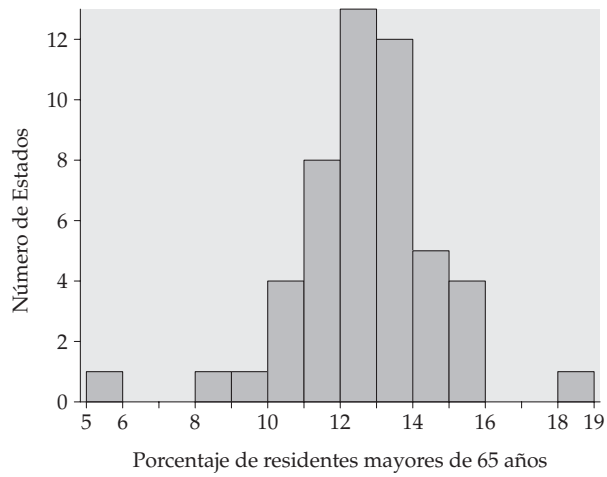

Figura 1.2. Histograma del porcentaje de residentes mayores de 65 años en los 50 Estados de EE UU. Datos de la tabla 1.1.

Las barras de un histograma deben cubrir todo el recorrido de una variable. Cuando haya saltos entre los posibles valores de la variable, extiende la base de las barras hasta llegar a medio camino de dos valores adyacentes posibles. Por ejemplo, en un histograma que muestra la edad 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.

Nuestra vista responde al *área* de las barras de un histograma.[1] Debido a que todas las clases tienen la misma anchura, el área está determinada por la altura y todas las clases se representan de forma equitativa. No hay una sola elección correcta del número de 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. 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.

Tabla 1.2. Consumos en carretera de coches de 1998.

Modelo	Consumo (litros/100 km)	Modelo	Consumo (litros/100 km)
Acura 3,5RL	9,5	Lexus GS300	10,3
Audi A6 Quattro	9,1	Lexus LS400	9,5
Buick Century	8,2	Lincoln Mark VIII	9,1
Cadillac Catera	9,9	Mazda 626	7,2
Cadillac Eldorado	9,1	Mercedes-Benz E320	8,2
Chevrolet Lumina	8,2	Mercedes-Benz E420	9,1
Chrysler Cirrus	7,9	Mitsubishi Diamante	9,9
Dodge Stratus	8,4	Nissan Maxima	8,4
Ford Taurus	8,4	Oldsmobile Aurora	9,1
Honda Accord	8,2	Rolls-Royce Silver Spur	14,8
Hyundai Sonata	8,5	Saab 900S	9,5
Infiniti I30	8,4	Toyota Camry	7,9
Infiniti Q45	10,3	Volvo S70	9,5

[1]Nuestros ojos responden al área, pero no de forma completamente lineal. Parece que percibimos la relación entre dos barras como el cociente entre las dos áreas elevado a 0,7. Véase W. S. Cleveland, *The Elements of Graphing Data*, Wadsworth, Monterey, Calif., 1985, págs. 278-284.

APLICA TUS CONOCIMIENTOS

1.5. Consumo de gasolina. El Ministerio de Industria exige que los fabricantes de automóviles den a conocer el consumo en ciudad y en carretera de cada modelo de automóvil. La tabla 1.2 muestra los consumos en carretera de 26 coches durante 1998.[2] Dibuja un histograma sobre los consumos en carretera de los automóviles.

1.2.3 Interpretación de los histogramas

Dibujar un gráfico estadístico no es un fin en sí mismo. Su objetivo es ayudarnos a comprender los datos. Después de hacer un gráfico, pregunta siempre: "¿qué veo?". Una vez hayas representado una distribución, puedes identificar sus características principales de la siguiente manera:

EXAMEN DE UNA DISTRIBUCIÓN

En cualquier gráfico de datos, identifica el **aspecto general** y las **desviaciones** sorprendentes del mismo.

Puedes describir el aspecto general de un histograma mediante su **forma**, su **centro** y su **dispersión**.

Un caso importante de desviación es una **observación atípica**, es decir, una observación individual que queda fuera del aspecto general.

En la sección 1.3 aprenderemos cómo describir numéricamente el centro y la dispersión. Por ahora, podemos describir el centro de una distribución mediante su *punto medio*, es decir, el valor tal que, de forma aproximada, la mitad de las observaciones son menores que él mismo y la otra mitad, mayores. Podemos describir la dispersión de una distribución dando los valores *mínimo* y *máximo*.

[2]U.S. Department of Energy, *Model Year 1998 Fuel Economy Guide*, Washington, D.C., 1997.

EJEMPLO 1.3. Descripción de una distribución

Fíjate otra vez en el histograma de la figura 1.2. **Forma**: la distribución es aproximadamente *simétrica* y tiene un *solo pico.* **Centro**: el punto medio de la distribución se halla próximo al pico, cerca del 13%. **Dispersión**: si ignoramos los cuatro valores más extremos, la dispersión va del 10 al 16%.

Observaciones atípicas: dos Estados se hallan en los extremos del histograma de la figura 1.2. Los puedes hallar en la tabla una vez el histograma te ha permitido identificarlos. Florida tiene un 18,5% de residentes de 65 o más años, mientras que Alaska tiene sólo un 5,2%. Una vez identificadas las observaciones atípicas, busca una explicación. Algunas observaciones atípicas se deben a errores, como por ejemplo escribir 50 en vez de 5,0. Otras observaciones atípicas indican la especial naturaleza de algunas observaciones. Florida, con mucha gente jubilada, tienen muchos residentes mayores de 65 años; en cambio, Alaska, en la frontera norte, tiene pocos. ■

Cuando describas una distribución, concéntrate en sus características principales. Fíjate en los picos mayores; no te preocupes por las pequeñas subidas y bajadas de las barras del histograma. Busca las observaciones atípicas claras; no busques sólo los valores máximo y mínimo. Identifica *simetrías* o *asimetrías* claras.

DISTRIBUCIONES SIMÉTRICAS Y 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. Una distribución **es asimétrica hacia la izquierda** si el lado izquierdo del histograma se extiende mucho más allá que el lado derecho.

En matemáticas, simetría significa que los dos lados de una figura, por ejemplo 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 los histogramas como el de la figura 1.2 son aproximadamente simétricos. Veamos más ejemplos.

EJEMPLO 1.4. Rayos en Colorado y Shakespeare

La figura 1.3 procede de un estudio sobre las tormentas acompañadas de aparato eléctrico en 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.

Por otro lado, la figura 1.4 muestra la distribución de la longitud de las palabras utilizadas en las obras de Shakespeare.[3] Esta distribución también tiene un solo pico, pero 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. ■

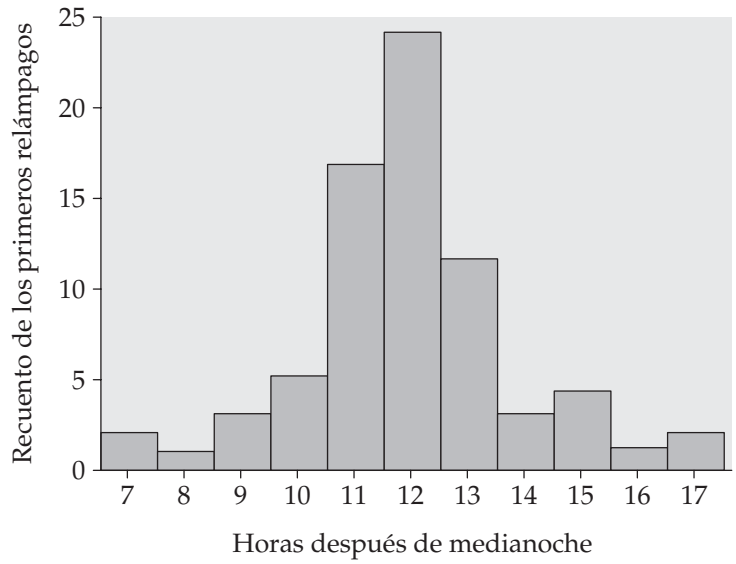

Figura 1.3. Distribución de la hora en la que se produce el primer relámpago del día en una localidad de Colorado, EE UU.

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

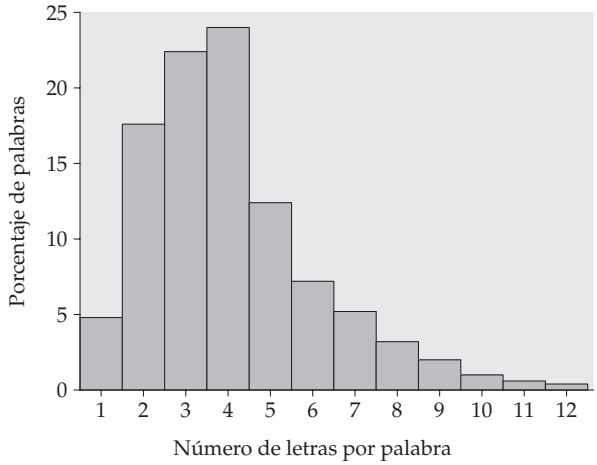

Figura 1.4. Distribución de la longitud de las palabras utilizadas en las obras de Shakespeare.

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.

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, a menudo, muy asimétricos hacia la derecha: hay muchos ingresos moderados, algunos elevados y muy pocos ingresos muy elevados. Recuerda que muchas distribuciones tienen formas que no pueden calificarse ni de simétricas ni de asimétricas. 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. Utiliza la vista y di lo que observas.

APLICA TUS CONOCIMIENTOS

1.6. Consumo de gasolina de automóviles. La tabla 1.2 proporciona datos sobre el consumo de automóviles. Basándote en el histograma de estos datos:

(a) Describe las características principales (forma, centro, dispersión y observaciones atípicas) de la distribución del consumo en carretera.

(b) El Gobierno impone un impuesto especial para coches con un consumo muy elevado. ¿Qué modelos crees que podrían ser objeto de este impuesto?

1.7. ¿Cómo describirías el centro y la dispersión de la distribución del primer relámpago del día de la figura 1.3? ¿Y de la distribución de la longitud de las palabras de la figura 1.4?

1.8. Rendimiento de acciones. El rendimiento total de una acción se obtiene teniendo en cuenta su precio de venta en Bolsa y los dividendos pagados por la empresa. El rendimiento total se expresa normalmente como un porcentaje sobre el precio de compra inicial. La figura 1.5 es un histograma sobre la distribución de los rendimientos totales de 1.528 acciones en la Bolsa de Nueva York durante un año.[4] 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.

(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 lo ha perdido?

1.2.4 Variables cuantitativas: 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.

[4]John K. Ford, "Diversification: how many stocks will suffice?" *American Association of Individual Investors Journal*, enero 1990, págs. 14-16.

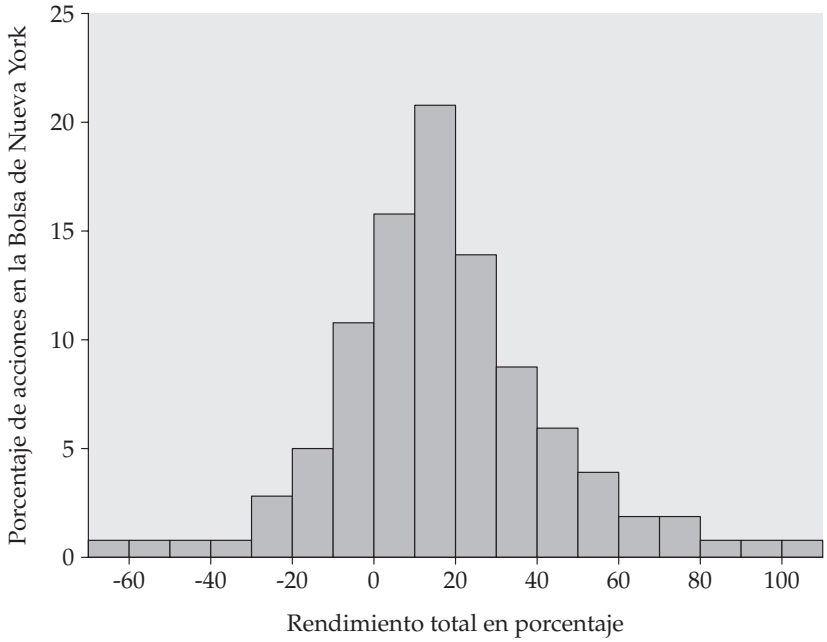

Figura 1.5. Distribución de rendimientos totales de todas las acciones de la Bolsa de Nueva York durante un año. Para el ejercicio 1.8.

DIAGRAMAS DE TALLOS

Para hacer un **diagrama de tallos**:

1. 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**, con el dígito del final. Los tallos pueden tener tantos dígitos como se quiera, pero cada hoja contiene un solo dígito.

2. Sitúa los tallos de forma vertical en orden creciente de arriba abajo. Traza una línea vertical a la derecha de los tallos.

3. Sitúa cada hoja a la derecha de su tallo, en orden creciente desde cada tallo.

```
 5 | 2
 6 |
 7 |
 8 | 8
 9 | 9
10 | 2 5
11 | 0 0 2 2 4 4 4 4 6
12 | 0 1 1 3 4 4 5 5 5 6 6 8 9
13 | 0 2 2 3 4 4 4 5 7 8 8 8 9
14 | 1 3 4 4 5
15 | 2 2 8 9
16 |
17 |
18 | 5
```

Figura 1.6. Diagrama de tallos correspondiente al porcentaje de residentes de 65 o más años en los Estados de EE UU. Compara este diagrama con el histograma de la figura 1.2.

EJEMPLO 1.5. Diagrama de tallos para los datos de "mayores de 65 años"

Para los porcentajes de "mayores de 65" 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, 13,0, tiene 13 de tallo y 0 de hoja. Los tallos pueden tener tantos dígitos como se necesiten, pero cada hoja tiene que consistir en un solo dígito. La figura 1.6 representa el diagrama de tallos correspondiente a los datos de la tabla 1.1.

Un diagrama de tallos tiene un aspecto parecido al de un histograma colocado en posición vertical. El diagrama de tallos de la figura 1.6 se parece al histograma de la figura 1.2. Los dos gráficos son ligeramente distintos debido a que las clases escogidas para el histograma no son iguales a los tallos del diagrama de tallos. 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. ■

En un histograma puedes escoger las clases. En cambio, las clases (los tallos) de un diagrama de tallos te vienen dadas. Puedes tener más flexibilidad **redondeando** los datos de manera que el dígito final, después del redondeo, sea

Redondeo

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. Debes redondear los datos como

3,5 2,6 3,0 1,1 ...

antes de dibujar el diagrama de tallos.

División de 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 la figura 1.6, por ejemplo, los tallos 12 y 13 serán

12	011344
12	5556689
13	0223444
13	578889

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 de la figura 1.6 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.

APLICA TUS CONOCIMIENTOS

1.9. Motivación y actitud de los estudiantes. 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)?

1.2.5 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 mostrarlos dibujaremos *un gráfico temporal.*

GRÁFICOS TEMPORALES

Un **gráfico temporal** de una variable representa cada observación en relación con el momento en que se midió. Sitúa siempre el tiempo en el eje de las abscisas. La unión de los puntos contiguos mediante segmentos facilita la visualización de los cambios a lo largo del tiempo.

EJEMPLO 1.6. Mortalidad por cáncer

He aquí los 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 1945 hasta 1995.

Año	1945	1950	1955	1960	1965	1970	1975	1980	1985	1990	1995
Muertos	134,0	139,8	146,5	149,2	153,5	162,8	169,7	183,9	193,3	203,2	204,7

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. De hecho, si ajustamos los datos de acuerdo con el incremento de edad de la población de EE UU, podemos ver que la tasa de muertes por cáncer ha ido disminuyendo desde 1992. ■

Cuando examines un gráfico temporal, fíjate una vez más en su aspecto general y en las desviaciones importantes de dicho aspecto. Un aspecto general que aparece con frecuencia 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.

Tendencia

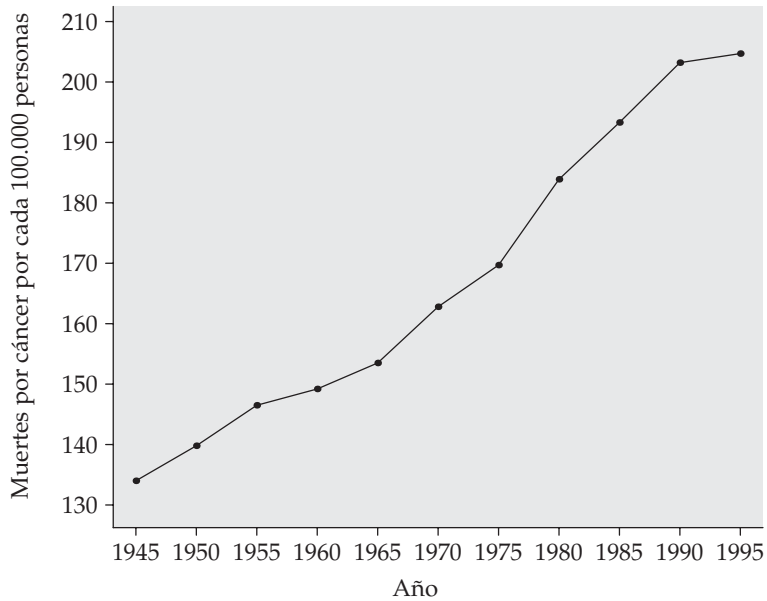

Figura 1.7. Gráfico temporal correspondiente a las tasas de mortalidad por cáncer en EE UU (número de muertes por cada 100.000 personas), desde 1945 hasta 1995.

APLICA TUS CONOCIMIENTOS

1.10. Fondos de inversión. Los intereses medios anuales, en porcentaje, pagados por unos determinados fondos de inversión en EE UU son los siguientes:[5]

Año	Intereses	Año	Intereses	Año	Intereses	Año	Intereses
1973	7,60	1979	10,92	1985	7,77	1991	5,70
1974	10,79	1980	12,88	1986	6,30	1992	3,31
1975	6,39	1981	17,16	1987	6,17	1993	2,62
1976	5,11	1982	12,55	1988	7,09	1994	3,65
1977	4,92	1983	8,69	1989	8,85	1995	5,37
1978	7,25	1984	10,21	1990	7,81	1996	4,80

[5]Albert J. Fredman, "A closer look at money market funds", *American Association of Individual Investors Journal*, febrero 1997, págs. 22-27.

(a) Dibuja un diagrama temporal con los intereses de los fondos de inversión.

(b) Las tasas de interés, al igual que muchas variables económicas, muestran **ciclos**, es decir, subidas y bajadas de su valor que aunque irregulares son claras. ¿En qué años aparecen picos temporales en los ciclos de la tasa de interés? *Ciclos*

(c) Además de la presencia de ciclos, los diagramas temporales pueden mostrar una tendencia consistente. De los años considerados, ¿en cuál se llega a alcanzar el valor máximo? A partir de ese año, ¿se observa una tendencia general decreciente?

RESUMEN DE LA SECCIÓN 1.2

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 alguna característica de un individuo, como puede ser la altura, el sexo o el salario.

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 euros.

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

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 **diagramas de barras** y los **diagramas de sectores** describen la distribución de variables categóricas. 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 del mismo.

La **forma**, el **centro** y la **dispersión** 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 correspondan a diferentes momentos del tiempo, haz un **gráfico temporal** situando la escala temporal en el eje de las abscisas 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

1.11. Salarios de técnicos de la FAO. He aquí una pequeña parte de un conjunto de datos que describe los salarios pagados por la Organización de las Naciones Unidas para la Agricultura y la Alimentación (FAO) a sus técnicos de alto nivel durante el periodo 1999/2000:

Técnico	Nacionalidad	Posición	Edad	Salario
Josep Ferre	Española	Oficial de enlace	38	58.378
Akima Mohamed	Marroquí	Coordinadora de programa	27	63.477
Robert Plumb	Británica	Oficial superior de finanzas	63	65.321
Jorge Pérez	Mexicana	Especialista en gestión	43	57.567

(a) ¿Qué individuos describe este conjunto de datos?

(b) Aparte del nombre de los técnicos, ¿cuántas variables contiene el conjunto de datos? De estas variables, ¿cuáles son categóricas y cuáles cuantitativas?

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

1.12. ¿A qué edad muere la gente joven? En 1997 las muertes de personas entre 15 y 24 años en EE UU se debieron a siete causas principales: accidentes, 12.958; homicidios, 5.793; suicidios, 4.146; cáncer, 1.583; enfermedades del corazón, 1.013; defectos congénitos, 383; y sida, 276.[6]

(a) Dibuja un diagrama de barras para mostrar la distribución de estos datos.

(b) Para dibujar un diagrama de sectores, ¿qué otra información necesitas?

1.13. Estilo de escritura y estadística. Los datos numéricos pueden distinguir diferentes estilos de escritura e incluso a veces hasta autores individuales. Tenemos datos sobre el porcentaje de palabras de 1 a 15 letras utilizadas en los artículos de la revista *Popular Science*:[7]

[6] *Births and Deaths: Preliminary Data for 1997*, Monthly Vital Statistics Reports, 47, nº 4, 1998.

[7] Datos obtenidos por estudiantes.